

# Content-aware Token Sharing for Efficient Semantic Segmentation with Vision Transformers

Chenyang Lu\*, Daan de Geus\*, Gijs Dubbelman  
Eindhoven University of Technology  
{c.lu.2, d.c.d.geus, g.dubbelman}@tue.nl

## Abstract

This paper introduces *Content-aware Token Sharing (CTS)*, a token reduction approach that improves the computational efficiency of semantic segmentation networks that use Vision Transformers (ViTs). Existing works have proposed token reduction approaches to improve the efficiency of ViT-based image classification networks, but these methods are not directly applicable to semantic segmentation, which we address in this work. We observe that, for semantic segmentation, multiple image patches can share a token if they contain the same semantic class, as they contain redundant information. Our approach leverages this by employing an efficient, class-agnostic policy network that predicts if image patches contain the same semantic class, and lets them share a token if they do. With experiments, we explore the critical design choices of CTS and show its effectiveness on the ADE20K, Pascal Context and Cityscapes datasets, various ViT backbones, and different segmentation decoders. With *Content-aware Token Sharing*, we are able to reduce the number of processed tokens by up to 44%, without diminishing the segmentation quality.

## 1. Introduction

In recent years, many works have proposed replacing Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) [13] to solve various computer vision tasks, such as image classification [1, 13, 22, 23], object detection [3, 23, 56] and semantic segmentation [6, 23, 28, 37, 47, 54]. ViTs<sup>1</sup> apply multiple layers of multi-head self-attention [41] to a set of tokens generated from fixed-size image patches. ViT-based models now achieve state-of-the-art results for various tasks, and it is found that ViTs are especially well-suited for pre-training on large datasets [1, 16], which in turn yields significant improvements on down-

\*Both authors contributed equally.

<sup>1</sup>In this work we use the term *ViTs* for the complete family of vision transformers that purely apply global self-attention.

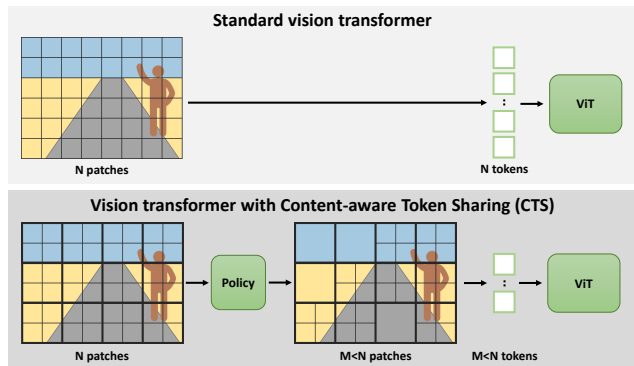


Figure 1. **Content-aware token sharing (CTS)**. Standard ViT-based segmentation networks turn fixed-size patches into tokens, and process all of them. To improve efficiency, we propose to let semantically similar patches share a token, and achieve considerable efficiency boosts without decreasing the segmentation quality.

stream tasks. However, the use of global self-attention, which is key in achieving good results, means that the computational complexity of the model is quadratic with respect to the input tokens. As a result, these models become particularly inefficient for dense tasks like semantic segmentation, which are typically applied to larger images than for image classification.

Recent works address these efficiency concerns in two different ways. Some works propose new ViT-based architectures to improve efficiency, which either make a combination of global and local attention [23, 48] or introduce a pyramid-like structure inspired by CNNs [14, 23, 44]. Alternatively, to reduce the burden of the quadratic complexity, some works aim to reduce number of tokens that are processed by the network, by either discarding [29, 39, 49] or merging [2, 21, 30] the least relevant tokens. These works, which all address the image classification task, find that a similar accuracy can be achieved when taking into account only a subset of all tokens, thereby improving efficiency. However, these methods, which are discussed further in Section 2, are not directly applicable to semantic segmentation. First, we cannot simply discard certain tokens, as each

token represents an image region for which the semantic segmentation task requires predictions. Second, existing token merging approaches allow any combination of tokens to be merged, through multiple stages of the network. As a result, ‘unmerging’ and spatial reorganization of tokens, which is necessary because semantic segmentation requires a prediction for each original token, is non-trivial.

In this work, we present a much simpler, yet highly effective and generally applicable token reduction approach for ViT-based semantic segmentation networks. The goal of this approach is to improve the efficiency without decreasing the segmentation quality. Token reduction methods for image classification already found that merging redundant tokens only results in a limited accuracy drop [2, 21, 30]. For semantic segmentation, we observe that there are many neighboring patches that contain the same semantic class, and that the information contained by these patches is likely redundant. Therefore, we hypothesize that neighboring patches containing the same class can share a token without negatively impacting the segmentation quality of a semantic segmentation network. To leverage this, we propose an approach that reduces tokens by (1) using a policy that identifies which patches can share a token before any tokens enter the ViT, and (2) grouping only rectangular neighboring regions, which allows for straightforward reassembling of tokens at the output of the ViT backbone. The main challenge that we tackle is to find this policy without introducing a heavy computational burden.

We propose to tackle this policy challenge by turning the problem into a simple binary classification task that can be solved by a highly efficient CNN model, our *policy network*. Concretely, it predicts whether  $2 \times 2$  neighboring patches contain the same class, and lets these patches share a token if they do. Our complete approach, which we call *Content-aware Token Sharing* (CTS), first applies this policy network, then lets patches share tokens according to the predicted policy, feeds the reduced set of tokens through the ViT, and finally makes a semantic segmentation prediction using these tokens (see Section 3). CTS has several advantages by design. First, it does not require any modifications to the ViT architecture or its training strategy for pre-training or fine-tuning. As a result, it is compatible with all backbones that purely use global self-attention, as well as any semantic segmentation decoder or advanced pre-training strategy, *e.g.*, BEiT [1]. Second, by reducing the number of tokens before inputting them to the ViT, the efficiency improvement is larger than when token reduction is applied gradually in successive stages of the ViT, as done in existing methods for image classification [2, 49]. Both advantages are facilitated by our policy network, which is trained separately from the ViT, and is so efficient that it only introduces a marginal computational overhead for large ViTs applied to high-resolution images.

With experiments detailed in Section 4, we show that our CTS approach can reduce the total number of processed tokens by at least 30% without decreasing the segmentation quality, for multiple datasets. We also show that this holds for transformer backbones of many different sizes, initialized with different pre-trained weights, and using various segmentation decoders. For more detailed results, we refer to Section 5. With this work, we aim to provide a foundation for future research on efficient transformer architectures for per-pixel prediction tasks. The code for this work is available through <https://tue-mps.github.io/CTS>.

To summarize, the contributions of this work are:

- A generally applicable token sharing framework that lets semantically similar neighboring image patches share a token, improving the efficiency of ViT-based semantic segmentation networks without reducing the segmentation quality.
- A content-aware token sharing policy network that efficiently classifies whether a set of neighboring patches should share a token or not.
- A comprehensive set of experiments with which we show the effectiveness of the proposed CTS approach on the ADE20K [55], Pascal Context [27], and Cityscapes [10] datasets, for a wide range of different ViTs and decoders.

## 2. Related work

**Efficient vision transformers.** Several works aim to mitigate the effect of the quadratic complexity of vision transformers, and make them more efficient. As mentioned in the introduction, different strategies exist to improve the efficiency of ViTs [14, 23, 44, 48], and our CTS falls in the category of approaches that aim to reduce tokens. This can be achieved by either discarding [29, 39, 49] or merging tokens [2, 21, 30, 33] at different stages of the network, and the number of processed tokens per image frequently depends on the complexity of the image. Each of these methods, which are strictly proposed for image classification, makes the decision to discard or merge tokens in a different way, *e.g.*, based on the similarity of tokens in embedding space [2], the attentiveness of tokens [21] or with learned parameters [30]. To be even more adaptive, AdaViT [26] and V-MoE [31] also introduce sparsity in the network architecture, by allowing tokens or entire images to be processed by a subset of the parameters, and DVT [45] varies the processed patch size based on the confidence of the network’s predictions. The findings of these works are similar: it is possible to reduce a significant number of tokens with little impact on the image classification accuracy, implying that many tokens are redundant.

**Semantic segmentation with vision transformers.** Vision transformers are also commonly applied to semantic segmentation, usually in one of two different ways. First,

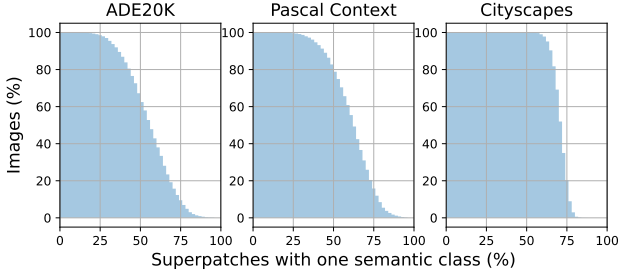


Figure 2. **Dataset statistics.** We show how many images have a certain percentage of superpatches that contain a single semantic class. We hypothesize that these superpatches can share a token.

there are works that propose custom transformer architectures that solve the entire task, for semantic segmentation [43, 47, 54] but also for more complex tasks like panoptic segmentation [18, 20, 42] and part-aware panoptic segmentation [11, 19]. Second, more commonly, works propose improvements to either a transformer-based backbone [8, 12, 15, 23, 44] or task-specific decoder [6, 7, 28, 37, 52]. An advantage of a modular backbone–decoder architecture is that backbones can simply be replaced by ones that are pre-trained on more data. This is important because it has been shown that pre-training transformers on large – sometimes even multimodal – datasets [1, 5, 16, 23, 36] leads to significant improvements on downstream tasks such as semantic segmentation. Being compatible with such advanced pre-training methods is therefore an important property of our CTS.

Several of the proposed works also focus on improving efficiency, usually by proposing architecture changes [23, 25, 43, 44, 51]. However, to the best of our knowledge, the findings from the field of image classification that certain tokens are redundant, have not been applied to semantic segmentation. In this work, inspired by these findings, we propose an approach that allows neighboring image patches to *share* a token when they have similar content, to improve the efficiency while maintaining the segmentation quality.

**Efficient semantic segmentation.** Many works focus purely on efficient semantic segmentation [25, 50, 51, 53]. However, these works generally propose architectures that are purely optimized for efficiency, and achieve a segmentation quality that is below the state of the art. In contrast, our approach is designed to improve the efficiency of any existing or future state-of-the-art ViT-based segmentation network, without compromising the segmentation quality.

### 3. Content-aware token sharing

#### 3.1. Problem definition

In ViT-based segmentation architectures, an input RGB image  $I \in \mathbb{R}^{3 \times H \times W}$  is partitioned into a set of  $N$  image patches  $\mathcal{I} = \{I_1, \dots, I_N\}$ . These image patches are linearly projected into a set of tokens  $\mathcal{T} = \{T_1, \dots, T_N\}$ , added to

their corresponding learnable positional embeddings, and are subsequently fed into transformer blocks [13]. The total number of tokens  $N$  depends on the size of the input image and the pre-defined patch size. A transformer-based backbone  $f : \mathcal{T} \mapsto \mathcal{L}$  maps input tokens  $\mathcal{T}$  into a set of predicted tokens  $\mathcal{L} = \{L_1, \dots, L_N\}$  containing semantic information. The transformer model can have any architecture applying purely global self-attention, *e.g.*, [1, 13]. The next step is to feed the output tokens  $\mathcal{L}$  into a decoder  $g$ . There are two options for this decoder. (1) If the decoder requires tokens, *e.g.*, a transformer decoder [37], the tokens are immediately fed through the decoder  $g : \mathcal{L} \mapsto \mathcal{O}_{tkn}$ , where  $\mathcal{O}_{tkn} = \{O_1, \dots, O_N\}$  is a set of per-token semantic predictions. These predictions are then rearranged and upsampled to per-pixel segmentation predictions  $O \in \mathbb{R}^{C \times H \times W}$  for  $C$  classes. (2) If the decoder requires spatially organized features, *e.g.*, a CNN-based segmentation head [46], the tokens are first rearranged based on their original position to get features  $L_{spat} \in \mathbb{R}^{E \times H_L \times W_L}$ , where  $E$  is the embedding dimension and  $H_L, W_L$  are fractions of the original  $H, W$ . These are then fed through the segmentation decoder  $g : L_{spat} \mapsto O$  to obtain the segmentation predictions.

**The problem** that is tackled by our approach is decreasing the size of  $\mathcal{T}$  to reduce computation, by identifying patches in  $\mathcal{I}$  that can share a token, without diminishing the segmentation quality of  $O = g(f(\mathcal{T}))$ .

#### 3.2. Content-aware token sharing framework

Our solution is based on the observation that some patches contain redundant information. In Figure 2, we show for several datasets the number of images that have a specific percentage of  $2 \times 2$  neighboring image patches, *i.e.*, *superpatches*, in  $\mathcal{I}$  that contain only one semantic class. We observe that in all images from all these datasets, roughly 25% of superpatches contain only one semantic class, and for many images this percentage is even higher. From these statistics, it is clear that there are many semantically uniform regions, and we hypothesize that efficiency can be improved without diminishing segmentation quality by processing these patches jointly. To achieve this, we propose *Content-aware Token Sharing* (CTS), which first predicts what superpatches in  $\mathcal{I}$  contain only one semantic class, and then lets each of these sets of  $2 \times 2$  neighboring patches share the same, single token.

To enable CTS for any conventional transformer-based model, we introduce a *token sharing* function  $t_s$ , a *token unsharing* function  $t_u$ , and a *policy* model  $p$ . All these are implemented in a differentiable manner, such that end-to-end training can be performed as usual.

The token sharing function  $t_s$  maps the original image patches  $\mathcal{I}$  and their positional embeddings into a new set of patches  $\mathcal{I}' = \{I'_1, \dots, I'_M\}$  with  $M < N$  samples, with corresponding positional embeddings. This mapping is per-

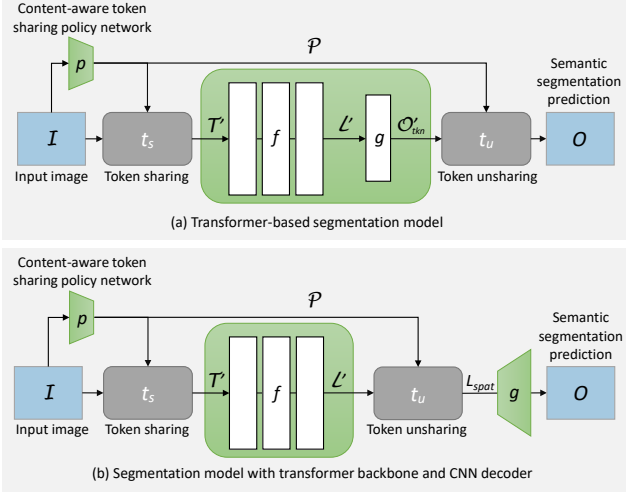


Figure 3. **Method overview.** We introduce a policy network  $p$  that predicts what image patches can share a token without decreasing the performance. These patches are then combined into a single token using token sharing module  $t_s$ . Subsequently, remaining tokens are fed through the transformer model, and the output tokens are ‘unshared’ using module  $t_u$ , either (a) after the per-token predictions are made or (b) before the per-pixel predictions are made.

formed according to a predicted policy  $\mathcal{P}$  that states which superpatches should share a single token, and which should have a unique token for each of their  $2 \times 2$  patches, as conventional. This policy is predicted by the policy model  $p$ , which is explained in detail in Section 3.3. The reduced set of image patches  $\mathcal{T}'$  is then projected to a reduced set of tokens  $\mathcal{T}'$  to which conventional transformer-based backbones can be deployed without any modifications. This token sharing function can be formalized as  $\mathcal{T}' = t_s(\mathcal{I}, \mathcal{P})$ . In this function, each superpatch that can share a token according to policy  $\mathcal{P}$  is bilinearly downsampled to the dimensions of a single patch, and linearly projected into a single token. As a result, the  $2 \times 2$  patches in this superpatch share the same token.

Given  $\mathcal{T}'$ , the transformer-based backbone  $f$  produces a set  $\mathcal{L}' = \{L'_1, \dots, L'_M\}$  of  $M$  predicted tokens with semantic information. Then, again, there are two options for the decoder. (1) When the set  $\mathcal{L}'$  including shared tokens is directly fed to the decoder, the resulting per-token predictions  $\mathcal{O}'_{tkn}$  need to be ‘unshared’. We use the policy  $\mathcal{P}$  to identify the predictions that are made for shared tokens, simply bilinearly upsample them and divide them back into individual predictions, before reorganizing all tokens spatially and upsampling them to yield the complete semantic segmentation predictions  $O$ . This is formalized in Eq. 1 and visualized in Figure 3a. (2) When the tokens in  $\mathcal{L}'$  should first be rearranged to features  $L_{spat}$  before entering the decoder,  $t_u$  first unshares the shared tokens by bilinearly upsampling and rearranging them to produce  $\mathcal{L}$ , before organizing  $\mathcal{L}$  spatially as features  $L_{spat}$  and feeding this through the de-

coder  $g$ , which outputs segmentation predictions  $O$ . This is formalized in Eq. 2 and visualized in Figure 3b.

$$\begin{aligned}
 \mathcal{P} &= p(I); & \mathcal{P} &= p(I); \\
 \mathcal{T}' &= t_s(\mathcal{I}, \mathcal{P}); & \mathcal{T}' &= t_s(\mathcal{I}, \mathcal{P}); & (1) & & (2) \\
 \mathcal{O}'_{tkn} &= g(f(\mathcal{T}')); & \mathcal{L}' &= f(\mathcal{T}'); \\
 O &= t_u(\mathcal{O}'_{tkn}, \mathcal{P}). & O &= g(t_u(\mathcal{L}', \mathcal{P})).
 \end{aligned}$$

Our framework is compatible with many different backbones, decoders, and policy models. This flexibility, as well as the consistent efficiency gains and quality preservation, is demonstrated extensively in Section 5.

### 3.3. Content-aware token sharing policy

The token sharing function  $t_s$  that lets patches share tokens relies on the output  $\mathcal{P}$  of a content-aware token sharing policy model  $p$  that predicts whether a superpatch (containing  $2 \times 2$  neighboring patches) can share the same token without diminishing segmentation quality. Realizing such a CTS policy is crucial for our framework, but non-trivial. Given our hypothesis that patches containing the same semantic class do not require separate, unique tokens, we propose to solve this problem by training and using a highly efficient CNN model to predict if a superpatch contains a single semantic class, without explicitly predicting the classes. If a superpatch indeed contains a single class, its patches can share a single token. This class-agnostic model, defined as  $\mathcal{P} = p(I)$ , can be implemented with any lightweight network, and simply performs binary classification on each superpatch region in an input image. From the output of the network, the  $S$  highest-scoring superpatches are selected to share a token, to end up with  $M = N - 3S$  tokens that are processed by the transformer-based network.

Our class-agnostic policy model is trained using a ground truth that is conveniently generated from the available semantic segmentation annotations. Concretely, for each superpatch in an image, we analyze how many classes it contains in the semantic segmentation ground truth. If there is just a single semantic class, the ground truth for the policy model  $p$  is set to true, otherwise it is false. This is visualized in Figure 4. Using this ground truth, the policy model is trained with a standard cross-entropy loss.

It is possible to train our policy network in such a simple and efficient way because of the available annotations for this task. Unlike image classification and object detection annotations, which do not have sufficient granularity to directly determine what image regions are relevant for processing, semantic segmentation annotations provide labels for each pixel. This allows us to directly translate our hypothesis about redundancy of semantically similar patches into a binary ground truth per superpatch. Moreover, the fact that our policy model is class-agnostic is key to its efficiency. Because the binary task is significantly less complex



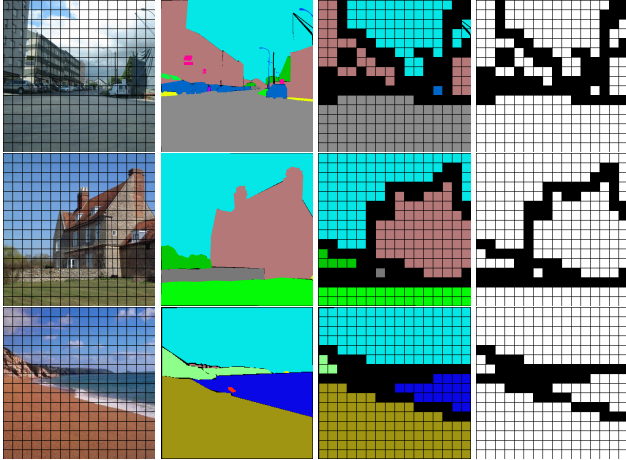


Figure 4. **Token sharing policy.** We teach our token sharing policy network that a superpatch should share a token if it contains a single semantic class. From left to right: (a) input image with grid of superpatches; (b) segmentation ground truth; (c) superpatches containing a single class; (d) class-agnostic policy ground truth.

than the main multi-class segmentation task, it can be performed by a highly efficient model, which is necessary to preserve the efficiency that is gained by reducing the number of tokens. In Section 5, we show that our class-agnostic approach outperforms a trivial solution that uses an auxiliary semantic segmentation prediction to obtain the semantically uniform superpatches.

## 4. Experimental setup

### 4.1. Datasets

To show the wide applicability of CTS for semantic segmentation, we conduct experiments on three datasets: ADE20K [55], Pascal Context [27], and Cityscapes [10]. Following most recent works, we use ADE20K for our main experiments and show the performance on the other two datasets to verify the general applicability of CTS.

### 4.2. Metrics

We use the conventional *mean intersection-over-union* (mIoU) to evaluate the segmentation performance. Since our work aims to improve the efficiency of segmentation networks, we also evaluate the throughput of each network, in terms of *images/second* (Im/sec). To measure this, we test each model on an Nvidia A100 GPU and report the average throughput over 100 iterations after 50 warm-up iterations, using batches of 32 image crops of the same resolutions used for training. For the proposed policy network, we additionally evaluate the *precision* of the predictions which superpatches should share a token. We choose precision because the ground truth can have a different number of shared superpatches for each image, while CTS always picks a fixed number of  $S$  superpatches to be shared.

### 4.3. Implementation details

**Policy network.** The content-aware token sharing policy network is trained separately from the semantic segmentation network. Our default policy network is an EfficientNet-Lite0 model [38], pre-trained on ImageNet-1K [32]. For all three datasets, we train the policy network for 50k iterations with batches of 8 images, using the AdamW [24] optimizer. The learning rate is  $5 \times 10^{-5}$  and weight decay is  $10^{-3}$ . During training, we apply the same resizing and cropping strategies used for the corresponding segmentation models.

**Segmentation networks.** Once the policy network is trained, we integrate it in the segmentation model and run the whole model end-to-end (see Figure 3). Specifically, we feed the training images through the policy network, apply the resulting policy to share tokens, and train the segmentation network on the resulting tokens. To train the segmentation networks, we follow the original training strategies, to allow for fair comparisons. For Segmenter [37], we optimize the networks with Stochastic Gradient Descent. For the ADE20K and Pascal Context datasets, the initial learning rate is set to 0.001, while for Cityscapes the initial learning rate is 0.01. All Segmenter experiments use polynomial learning rate decay, following the original settings [37]. The batch size is set to 8 for all the experiments. For experiments with the UPerNet decoder [46], we add our CTS to publicly available BEiT+UPerNet and ViT+UPerNet implementations [9] and use the original training and evaluation settings. All backbones except for BEiT are pre-trained on ImageNet-1K [32, 36]. For more details, we refer to the supplementary material.

## 5. Experimental results

### 5.1. Content-aware token sharing

To show the effectiveness of the proposed content-aware token sharing (CTS) approach, we apply it to ViT-based segmentation network Segmenter [37]. In Table 1, we show the results for different values of  $S$ , *i.e.*, different token sharing settings, leading to different token reductions and efficiency boosts. When  $S = 0$ , and no superpatches share a token, this is equal to the standard Segmenter [37]. When  $S = M = 256$ , all superpatches share a token, leading to a token reduction of 75%. From the results, we can see that reducing the number of tokens leads to an increase in throughput, as expected. Moreover, the results show that with CTS, up to 30% of tokens can be reduced without compromising the segmentation quality, with a mIoU of 45.1 compared to a baseline of 45.0. In a naive token reduction setting, however, when random superpatches are selected to be shared, the mIoU is significantly worse, *i.e.*, 42.9. This confirms our motivation that specific superpatches can share a token without diminishing the segmentation quality, if they are selected carefully, as with our CTS. And al-

Tokens			CTS		Random sharing	
Shared (S)	Total (M)	Reduction	mIoU	Im/sec	mIoU	Im/sec
0	1024	0%	45.0	122	-	-
31	931	9%	45.3	123	44.7	139
41	901	12%	44.9	126	44.5	143
79	787	23%	45.0	146	43.6	168
103	715	30%	45.1	162	42.9	191
131	631	38%	44.7	180	42.4	216
156	556	46%	44.4	198	42.5	242
192	448	56%	43.8	228	41.3	288
224	352	66%	41.5	267	40.8	354
256	256	75%	39.9	424	-	-

Table 1. **Content-aware token sharing.** The mean IoU and throughput of the Segementer model [37] with ViT-S/16 backbone [13] with our CTS and a random sharing approach for different token sharing settings. Results on the ADE20K val set [55].

though random token sharing is faster than our CTS, due to the overhead introduced by the policy network, CTS still achieves a significant throughput increase of 33% while keeping the segmentation quality the same.

With the identified optimal token reduction setting of 30% for ADE20K, we also apply our CTS approach to Segementer with larger and smaller backbones. From the results, depicted in Figure 5, it is immediately clear that CTS effectively improves the throughput for all backbone sizes by significantly reducing the number of tokens, while delivering a similar semantic segmentation performance. This further validates our approach.

## 5.2. Ablations

**Token sharing policy generation.** To analyze the role of the token sharing policy generation method  $p$  in CTS, we apply and compare different methods. These methods can be categorized into two groups: (1) using our policy network (PolicyNet) to directly predict the class-agnostic CTS policy, and (2) using semantic segmentation predictions to determine the CTS policy, *i.e.*, following the procedure in Figure 4 but with predicted semantic segmentation. The results are provided in Table 2. First of all, these results show that turning the problem into a class-agnostic classification task, instead of using the trivial solution of solving the full class-wise segmentation task, is necessary to achieve the best results. Moreover, this classification task allows us to solve it with a lightweight PolicyNet, which allows for a much higher throughput than conventional segmentation networks (162 vs. 60). Finally, we see that the performance can be boosted by having a larger PolicyNet, but at the expense of efficiency.

**Different backbones.** The CTS approach is compatible with any ViT-based backbone that purely applies global self-attention. In Table 3, we apply CTS to three backbones of similar sizes, which have different settings for weight initialization [1], distillation [40] and positional embeddings [1]. For all three backbones, our CTS approach

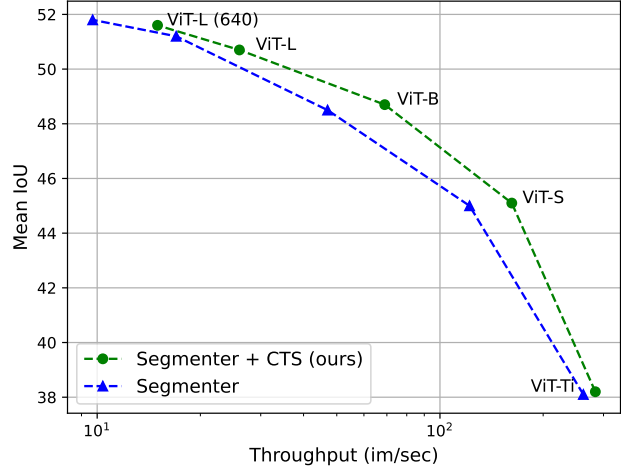


Figure 5. **Throughput vs. mean IoU.** The efficiency and accuracy of Segementer [37] with and without our content-aware token sharing approach. Results on the ADE20K val set [55].

Category	Policy method	Precision (policy)	Throughput (im/sec)	mIoU (seg.)
PolicyNet	PolicyNet (default)	80.3	162	45.1
	PolicyNet (large)	79.3	148	45.3
	PolicyNet (small)	73.9	186	44.0
Seg. pred.	MobileNetV2 + FCN [34, 35]	70.7	125	44.3
	DeepLabv3+ (RN-101) [4, 17]	74.2	60	44.8
	Ground-truth seg.	96.1	189	47.2
Random	Random	52.1	191	42.9
	None (baseline)	-	122	45.0

Table 2. **Different policy methods.** Ablation of policy generation methods applied to Segementer with ViT-S/16, on the ADE20K val set [55]. The token reduction schedule is set to 30%.

CTS	Backbone	Token reduction	Im/sec	mIoU
-	ViT-B/16 [13]	0%	47	48.5
✓		30%	69	48.7
-	DeiT-B/16 <sup>†</sup> [40]	0%	48	48.0
✓		30%	70	47.9
-	BEiT-B/16 [1]	0%	36	50.3
✓		30%	63	50.4

Table 3. **Different backbones.** The mean IoU and throughput of the Segementer model [37] with our CTS method, for different backbones. Results on the ADE20K val set [55]. <sup>†</sup>Distilled.

consistently results in an increase in throughput by reducing the number of tokens, without any loss in segmentation performance. This shows the general applicability of CTS.

**Different decoders.** To further show the general applicability of CTS, we also show that it is compatible with multiple decoders. We apply CTS to three different types of decoders to generate the semantic predictions: 1) *Linear*: a vanilla linear projection, 2) *Mask Transformer*: a

CTS	Decoder	Token reduction	Im/sec	mIoU
-	Linear	0%	57	48.1
✓		30%	82	47.9
-	Mask Transformer [37]	0%	47	48.5
✓		30%	69	48.7
-	UPerNet [46]	0%	21	48.1
✓		30%	28	47.7

Table 4. **Different decoders.** Our approach is compatible with multiple decoders, including transformer-based [37] or CNN-based [46] ones. Each network is trained with the ViT-B/16 backbone. Results on the ADE20K val set [55].

Method	Pre-training	Token reduction	Im/sec	mIoU (ms)
BEiT-L* + UPerNet	BEiT + IN22k	0%	3.3	56.8
BEiT-L* + UPerNet + CTS (ours)	BEiT + IN22k	30%	5.4	56.6
Swin-L + UPerNet [12,46]	IN22k	n/a		53.5
ViT-Adapter-L + UPerNet [5,46]	IN22k	n/a		54.4
SwinV2-L + UPerNet [6,22]	IN22k	n/a		55.9
Swin-L + Mask2Former [6,23]	IN22k	n/a		57.3
BEiT-L + UPerNet [1,46]	BEiT + IN22k	n/a		57.0
ViT-Adapter-L + UPerNet [5,46]	BEiT + IN22k	n/a		58.4
ViT-Adapter-L + Mask2Former [5]	BEiT + IN22k	n/a		59.0

Table 5. **Application to state-of-the-art.** We apply CTS to state-of-the-art method BEiT-L + UPerNet [1,46] and report other top-performing methods for reference. Results on the ADE20K [55] dataset, with multi-scale testing. \*Our implementation [9].

transformer-based decoder proposed in Segmenter [37] and deployed in our main experiments, and 3) *UPerNet*: a commonly-used CNN-based decoder [46]. Again, the results in Table 4 show that CTS considerably improves the throughput, *i.e.*, by up to 47%, while achieving a similar segmentation quality.

### 5.3. Application to state-of-the-art

We apply CTS to state-of-the-art method BEiT-L + UPerNet [1,46], and report the results in Table 5. For reference, we also show the results of other state-of-the-art networks with similar backbones and pre-training. The results show that our CTS significantly improves the throughput, by nearly 64%, while achieving a similar segmentation quality. This shows that CTS is also effective when applied to state-of-the-art methods, that it is compatible with large-scale self-supervised pre-training methods [1] that have shown to significantly boost the performance for downstream tasks, and that it is competitive with top-performing ViT-based segmentation networks.

### 5.4. Other datasets

To show that CTS is also effective on other datasets, we evaluate it on Pascal Context [27] and Cityscapes [10]. We apply CTS to Segmenter, report the results in Table 6, and compare to random token sharing for reference. From the results, it is clear that CTS consistently achieves a mean

Method	Backbone	Token reduction	Im/sec	mIoU
<i>Pascal Context validation</i>				
Segmenter	ViT-S/16	0%	157	53.0
Segmenter + CTS (rand.)	ViT-S/16	30%	246	51.3
Segmenter + CTS (ours)	ViT-S/16	30%	203	52.9
Segmenter	ViT-L/16	0%	21	57.7
Segmenter + CTS (ours)	ViT-L/16	30%	31	57.6
<i>Cityscapes val</i>				
Segmenter	ViT-S/16	0%	38	76.5
Segmenter + CTS (rand.)	ViT-S/16	44%	93	72.6
Segmenter + CTS (ours)	ViT-S/16	44%	78	76.5
Segmenter	ViT-L/16	0%	6.0	79.1
Segmenter + CTS (ours)	ViT-L/16	44%	12.6	79.5

Table 6. **Other datasets.** CTS with Segmenter [37] on the Pascal Context [27] and Cityscapes [10] datasets.

IoU similar to the standard Segmenter, while improving the throughput by up to 48% for Pascal Context and 110% for Cityscapes. The higher efficiency boost for Cityscapes is caused by a higher token reduction, which is possible because this dataset contains images in which more superpatches contain just a single class, as depicted in Figure 2, which means that more patches can share a token without diminishing the segmentation performance.

### 5.5. Additional analyses

**Consistency in predictions for shared tokens.** With our policy network, we aim to let superpatches share a token if they contain a single semantic class, and in Table 2 we have shown that the network mostly correctly predicts this. However, it is still unclear if the segmentation network subsequently also predicts that these superpatches contain a single semantic class. To check this, we enforce that each pixel in a superpatch that shares a token gets the same class prediction, and evaluate the impact on the mean IoU. We achieve this single-class prediction with a majority vote in each superpatch that shares a token. The results, reported in Table 7a, show that for token reductions up to 30%, the mean IoU remains constant. This indicates that there is little to no difference in the predictions, meaning that the network already makes consistent predictions within superpatches that share tokens. This further shows that our policy network is effective, and that the segmentation network automatically learns that these superpatches contain only one class, which is desired behavior.

**Dynamic token sharing.** The proposed CTS method uses a fixed number of  $S$  shared tokens for all the images in a dataset. However, images typically have different semantic complexities, even in the same dataset (see Figure 2). To evaluate whether it is desirable to let  $S$  depend on the complexity of the image, we conduct an experiment where we first apply our policy network to each image, threshold its prediction, and identify how many superpatches  $S^*$  can

Token reduction	mIoU		
	default	one class	diff.
12%	44.9	44.9	0.0
30%	45.1	45.1	0.0
56%	43.8	43.4	-0.4
66%	41.5	40.8	-0.7

(a) **Consistency in predictions for shared tokens.** We force all pixels in superpatches to have the same semantic class prediction if they share a token, and evaluate the mean IoU.

Token reduction	Processed images	mIoU
0%	2	45.0
12%	87	44.9
23%	191	45.0
30%	398	45.1
38%	427	44.7
46%	895	44.4
Combined	2000	<b>45.8</b>

(b) **Dynamic token sharing.** Results of our approach when images are assigned to models with different token sharing settings, depending on the predicted policy.

Using prev. frame	Policy	Precision (policy)	mIoU (seg.)
✓	Segmentation pred.	85.9	75.5
✓	PolicyNet	94.9	76.2
✗	PolicyNet (default)	96.8	76.5
✗	Random	69.5	72.6
✗	None (baseline)	-	76.5
✗	Ground truth	99.6	78.0

(c) **CTS on video data.** Experiments with video inputs on the Cityscapes val set [10]. The token reduction setting is 44%.

Table 7. **Additional analyses.** Networks trained with the ViT-S/16 backbone. All results on ADE20K val [55] unless indicated otherwise.

share a token according to this prediction. Subsequently, we pick the highest token sharing setting  $S$  for which  $S < S^*$ , let the  $S$  highest-scoring superpatches share a token, and feed the resulting tokens through a segmentation network trained with this setting  $S$ . Finally, we collect all segmentation predictions and calculate the mean IoU.

Table 7b presents the results for this experiment, together with the number of images assigned to each model. Interestingly, we find that the resulting mean IoU is 45.8, which is 0.7 points higher than the best individual model. This shows that, by dynamically determining how many tokens should be shared in each image, the segmentation quality can be even higher, because we do not force superpatches to share a token if the policy network assigns them a low score. This means that a future single-network, variable-token approach could optimize the accuracy and efficiency even further. Because the number of shared tokens in such a method would vary, there would be no guaranteed efficiency boost, introducing a new accuracy-efficiency trade-off.

**CTS on video data.** In the situation where there is a continual stream of images, *i.e.*, with video data, there is a large degree of consistency between subsequent frames. If this consistency is high enough, the predictions for a previous frame could be used to generate a policy for a current frame, removing the need for a separate policy network, which could improve the efficiency. We evaluate this for two settings: (1) we use the previous segmentation predictions to generate a policy according to the procedure in Figure 4, and (2) we run our policy network on the previous frame and use its output as the policy. The results, presented in Table 7c, show that using information from previous frames results in a segmentation quality only slightly below the original CTS approach, especially for the PolicyNet, *i.e.*, -0.3. Such an approach can be more efficient than the original policy network if the policy network is executed in parallel with the segmentation network, and it could be interesting for future research. Again, this experiment also shows the strength of our policy network, as it outperforms the trivial solution of using the previous semantic segmentation output as a policy.

## 6. Discussion

The experimental results demonstrate the positive impact of token sharing on the efficiency of ViT-based semantic segmentation networks. We show that the number of tokens can be reduced by 30% to 44% without compromising segmentation quality, depending on the dataset. Key to maintaining segmentation quality is selecting the correct superpatches for token sharing, *i.e.*, the superpatches containing a single semantic class. Our CTS approach achieves this with an efficient policy network that is conveniently trained in a supervised manner using the already available ground-truth data. Especially for large ViT models applied to high-resolution images, the computational overhead that our policy network introduces is marginal, and a significant improvement in throughput is achieved.

The strength of our CTS framework is its conceptual simplicity, which allows it to be very efficient and generally applicable, making it compatible with many vision transformers, including those using advanced pre-training. Such pre-training methods have shown to be key to obtaining state-of-the-art results [1] and our CTS enables such current and future methods to be significantly more efficient.

Due to its simple nature, CTS can be extended in several different ways in future work. First of all, it is conceptually compatible with alternative efficient approaches that combine global and local self-attention, such as Swin [23], and these methods have complementary advantages. Second, our token sharing approach is also compatible with other computation-intensive per-pixel tasks, such as optical flow and depth estimation. CTS can be applied to improve the efficiency on those tasks, but would require task-specific policy models. With this work, we lay the foundation for such follow-up research on efficient transformer-based architectures for per-pixel computer vision tasks.

**Acknowledgements** This work is supported by Eindhoven Engine, NXP Semiconductors and Brainport Eindhoven. This work made use of the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-3836, which is financed by the Dutch Research Council (NWO).



## References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *ICLR*, 2022. 1, 2, 3, 6, 7, 8
- [2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token Merging: Your ViT but Faster. In *ICLR*, 2023. 1, 2
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *ECCV*, 2020. 1
- [4] 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. 6
- [5] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *ICLR*, 2023. 3, 7
- [6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-Attention Mask Transformer for Universal Image Segmentation. In *CVPR*, 2022. 1, 3, 7
- [7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *NeurIPS*, 2021. 3
- [8] 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. 3
- [9] MMSegmentation Contributors. MMSegmentation: Open-MMLab Semantic Segmentation Toolbox and Benchmark. <https://github.com/open-mmlab/mms Segmentation>, 2020. 5, 7
- [10] 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 *CVPR*, 2016. 2, 5, 7, 8
- [11] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware Panoptic Segmentation. In *CVPR*, 2021. 3
- [12] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. CSWin Transformer: A General Vision Transformer Backbone With Cross-Shaped Windows. In *CVPR*, 2022. 3, 7
- [13] 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. In *ICLR*, 2021. 1, 3, 6
- [14] Benjamin Graham, Alaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Herve Jegou, and Matthijs Douze. LeViT: A Vision Transformer in ConvNet’s Clothing for Faster Inference. In *ICCV*, 2021. 1, 2
- [15] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z. Pan. Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation. In *CVPR*, 2022. 3
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 2022. 1, 3
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 6
- [18] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic Segmentation. In *CVPR*, 2019. 3
- [19] Xiangtai Li, Shilin Xu, Yibo Yang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Panoptic-PartFormer: Learning a Unified model for Panoptic Part Segmentation. In *ECCV*, 2022. 3
- [20] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo, and Tong Lu. Panoptic SegFormer: Delving Deeper Into Panoptic Segmentation With Transformers. In *CVPR*, 2022. 3
- [21] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In *ICLR*, 2022. 1, 2
- [22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *CVPR*, 2022. 1, 7
- [23] 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. In *ICCV*, 2021. 1, 2, 3, 7, 8
- [24] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 5
- [25] Sachin Mehta and Mohammad Rastegari. MobileViT: Lightweight, General-purpose, and Mobile-friendly Vision Transformer. In *ICLR*, 2022. 3
- [26] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. AdaViT: Adaptive Vision Transformers for Efficient Image Recognition. In *CVPR*, 2022. 2
- [27] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *CVPR*, 2014. 2, 5, 7
- [28] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *ICCV*, 2021. 1, 3
- [29] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In *NeurIPS*, 2021. 1, 2
- [30] Cedric Renggli, André Susano Pinto, Neil Houlsby, Basil Mustafa, Joan Puigcerver, and Carlos Riquelme. Learning to Merge Tokens in Vision Transformers. *arXiv preprint arXiv:2202.12015*, 2022. 1, 2
- [31] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel

- Keyzers, and Neil Houlsby. Scaling Vision with Sparse Mixture of Experts. In *NeurIPS*, pages 8583–8595, 2021. 2
- [32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 5
- [33] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive Space-Time Tokenization for Videos. In *NeurIPS*, 2021. 2
- [34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 2018. 6
- [35] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE TPAMI*, 39(4):640–651, 2017. 6
- [36] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *TMLR*, 2022. 3, 5
- [37] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for Semantic Segmentation. In *ICCV*, 2021. 1, 3, 5, 6, 7
- [38] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*, 2019. 5
- [39] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch Slimming for Efficient Vision Transformers. In *CVPR*, 2022. 1, 2
- [40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 6
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017. 1
- [42] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-End Panoptic Segmentation With Mask Transformers. In *CVPR*, 2021. 3
- [43] Jian Wang, Chenhui Gou, Qiman Wu, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. RTFormer: Efficient Design for Real-Time Semantic Segmentation with Transformer. In *NeurIPS*, 2022. 3
- [44] 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 *ICCV*, 2021. 1, 2, 3
- [45] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not All Images are Worth 16x16 Words: Dynamic Transformers for Efficient Image Recognition. In *NeurIPS*, 2021. 2
- [46] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *ECCV*, 2018. 3, 5, 7
- [47] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS*, 2021. 1, 3
- [48] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal Attention for Long-Range Interactions in Vision Transformers. In *NeurIPS*, 2021. 1, 2
- [49] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-ViT: Adaptive Tokens for Efficient Vision Transformer. In *CVPR*, 2022. 1, 2
- [50] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation. In *ECCV*, 2018. 3
- [51] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. TopFormer: Token Pyramid Transformer for Mobile Semantic Segmentation. In *CVPR*, 2022. 3
- [52] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-Net: Towards Unified Image Segmentation. In *NeurIPS*, 2021. 3
- [53] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In *ECCV*, 2018. 3
- [54] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xi Tian 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. 1, 3
- [55] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *CVPR*, 2017. 2, 5, 6, 7, 8
- [56] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR*, 2021. 1