Segmenter: Transformer for Semantic Segmentation

Robin Strudel* [1,2], Ricardo Garcia* [1,2], Ivan Laptev [1,2], Cordelia Schmid [1,2]

[1] INRIA Paris [2] DI ENS, PSL

2021 ICCV OCTOBER 11-17 VIRTUAL

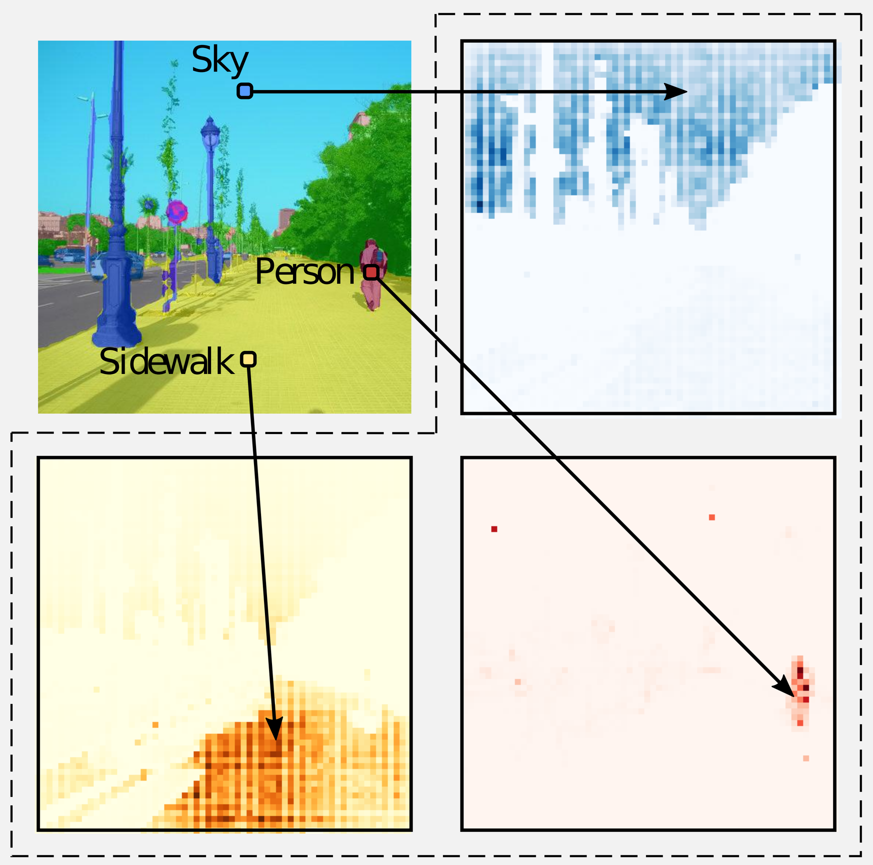
Motivation

- State-of-the-art methods deploy Fully Convolutional Networks (FCN) to achieve excellent results on Semantic Segmentation.

- The local nature of convolutional filters, however, results in features capturing only local context coming from neighbouring pixels.

- Global information is key to perform accurate segmentation as pixel level labeling often depends on the global image context.

- Transformer architectures can be used to leverage contextual information at every layer of the model:



Contributions

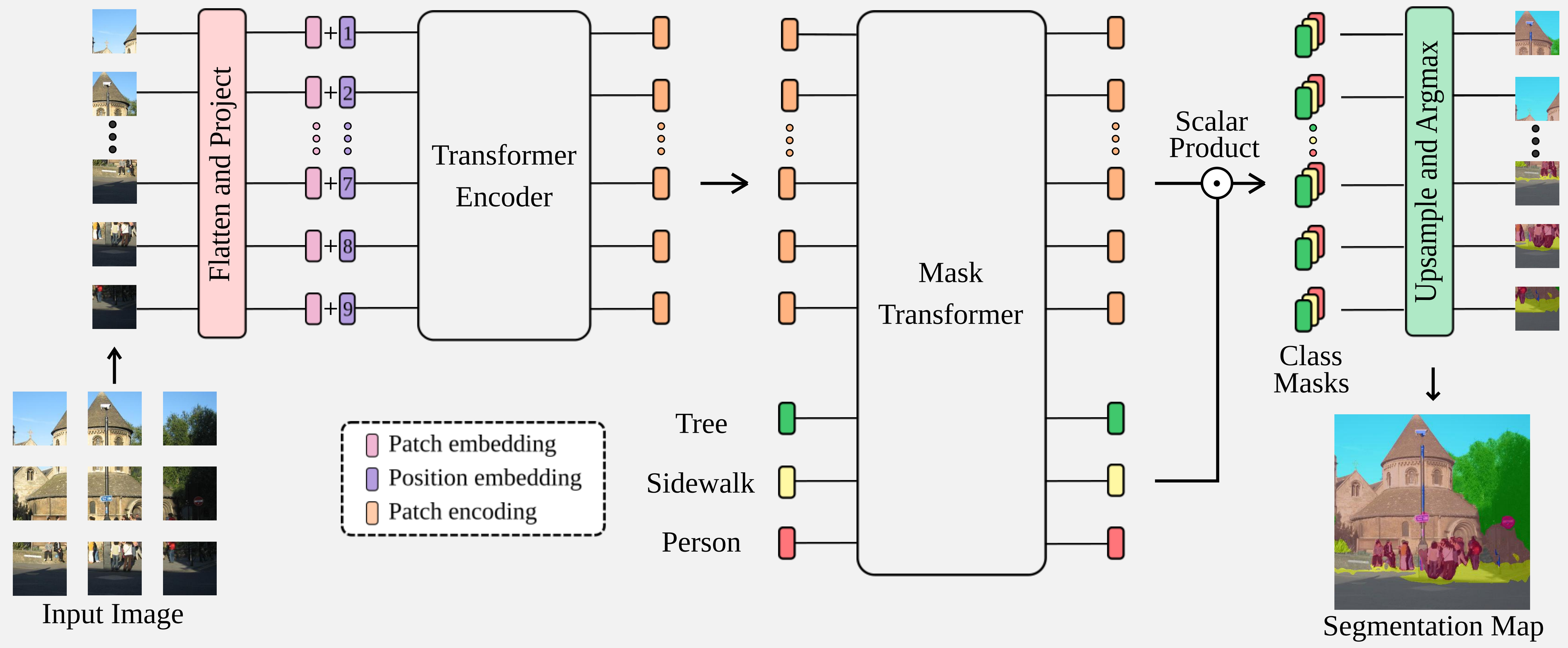
1. Novel approach to semantic segmentation based on ViT, capturing contextual information by design

2. Transformer-based decoder generating class masks for general image segmentation tasks.

3. State-of-the-art performance on ADE20K and Pascal Context.

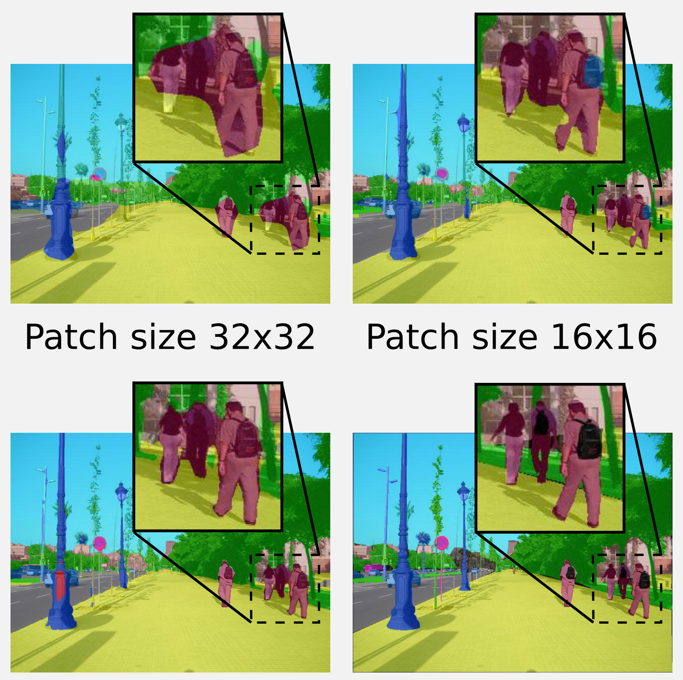
github.com/rstrudel/segmenter

Segmenter Architecture



(Left) Encoder: The image patches are projected to a sequence of embeddings and then encoded with a transformer. (Right) Decoder: A mask transformer takes as input the output of the encoder and class embeddings to predict segmentation masks.

Impact of Patch Size



Smaller patch size results in:

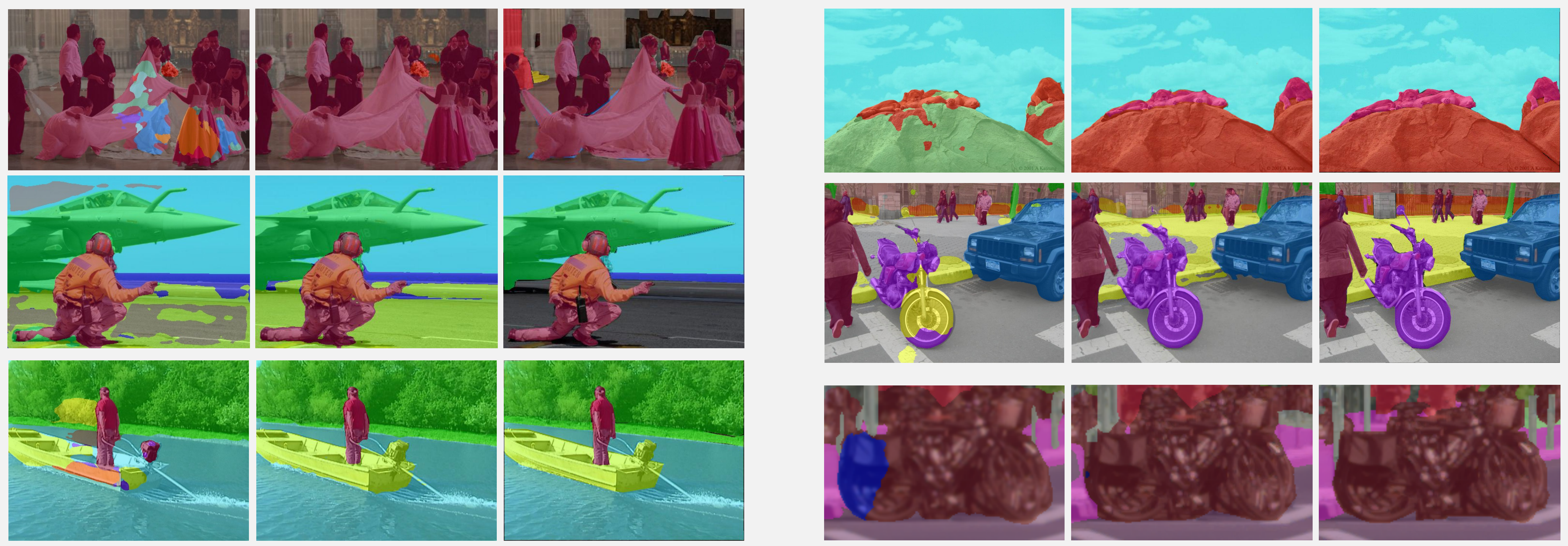
- Better boundaries in segmentation masks.

- Longer sequences, increasing the compute time and memory footprint.

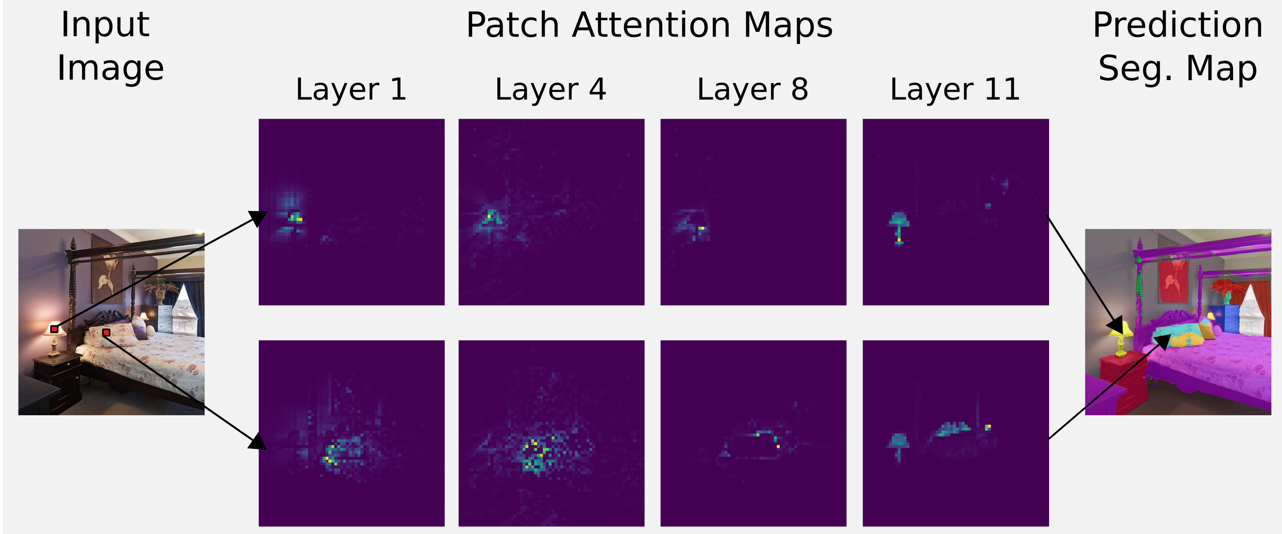
State-of-the-art ADE20K

Method	Backbone	Im/sec	mIoU	+MS
OCR [60]	HRNetV2-W48	83	-	45.66
ACNet [24]	ResNet-101	-	-	45.90
DNL [57]	ResNet-101	-	-	45.97
DRANet [22]	ResNet-101	-	-	46.18
CPNet [58]	ResNet-101	-	-	46.27
DeepLabv3+ [10]	ResNet-101	76	45.47	46.35
DeepLabv3+ [10]	ResNeSt-101	15	46.47	47.27
DeepLabv3+ [10]	ResNeSt-200	-	-	48.36
SETR-L MLA [67]	ViT-L/16	34	48.64	50.28
Swin-L UperNet [35]	Swin-L/16	34	52.10	**53.50**
Seg-B†/16	DeiT-B/16	77	47.08	48.05
Seg-B†-Mask/16	DeiT-B/16	76	48.70	50.08
Seg-L/16	ViT-L/16	33	50.71	52.25
Seg-L-Mask/16	ViT-L/16	31	51.82	**53.63**

Qualitative Results



Encoder: Patch Attention Maps

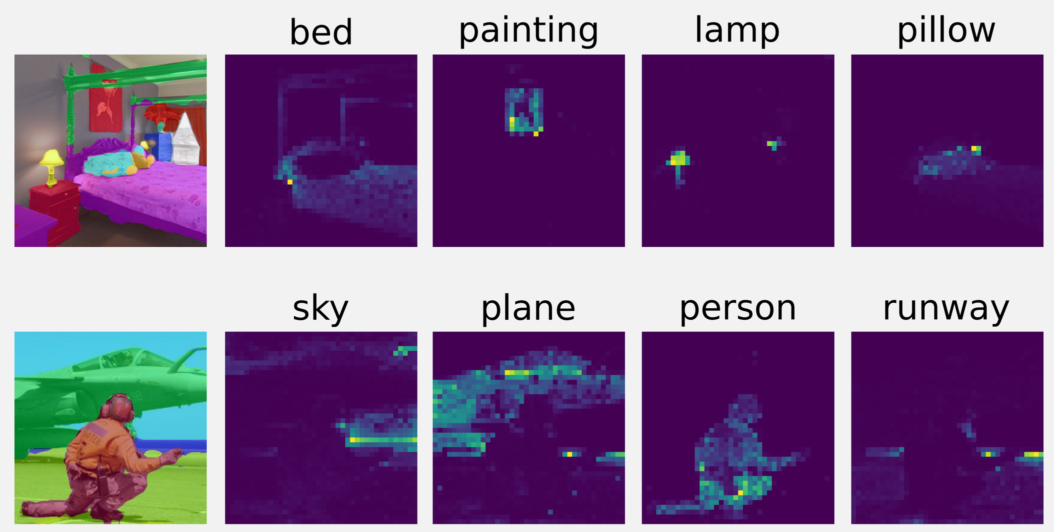


The attention map field-of-view adapts to the input image and instances size:

- Gathering global information on large instances, such as the bed.

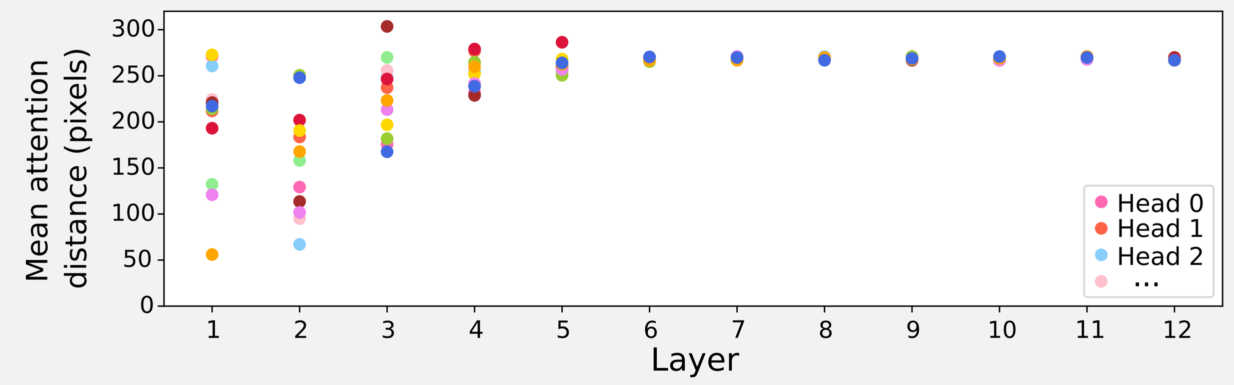
- Focusing on local information on smaller instances, such as the lamp.

Decoder: Class Attention Maps



Class embeddings from the mask transformer attends to patch embeddings corresponding to its class in the input image.

Size of Attended Area



Segmenter provides larger receptive field size than CNNs.

Already at the first layer, some heads attend to most of the image and distant patches which clearly lie outside the receptive field ResNet. initial layers.