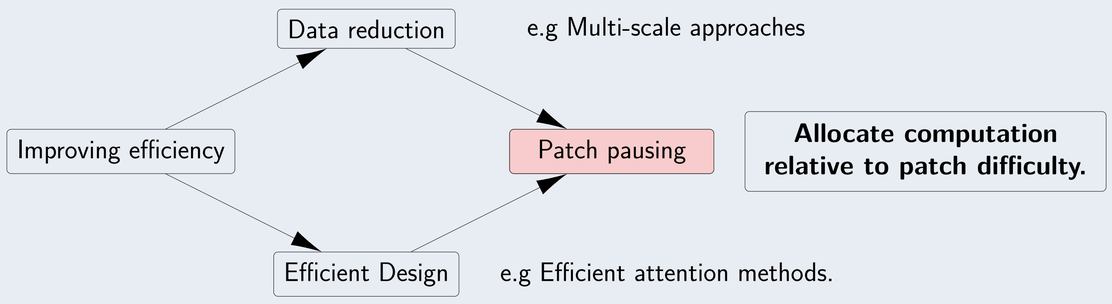
PAUMER: Patch Pausing Transformer for Semantic Segmentation

Evann Courdier [*,a,b], Prabhu Teja S [*,a,b] François Fleuret[a, b, c]

[*]Equal Contribution, [a]Idiap Research Institute, [b]EPFL, [c]University of Geneva, Switzerland,
Contact: evann.courdier@idiap.ch, prabhu.teja@idiap.ch

Efficient Segmentation by non-uniform allocation of computation



Reduce computation for 'easier' parts of the input image.

Not all tokens are equal

(a) Input image. (b) The first layer at which the patch is correctly classified. Darker is lower.

Figure: How long do we have to process each part of the image for correct predictions with a ViT-Ti backbone? It is apparent that not all patches of the images need a homogenous amount of processing i.e., not all patches need the full network and can be reliably decoded after a few layers itself. We exploit this to build patch pausing for efficient transformers for segmentation.

Pause processing tokens which have been processed 'enough'

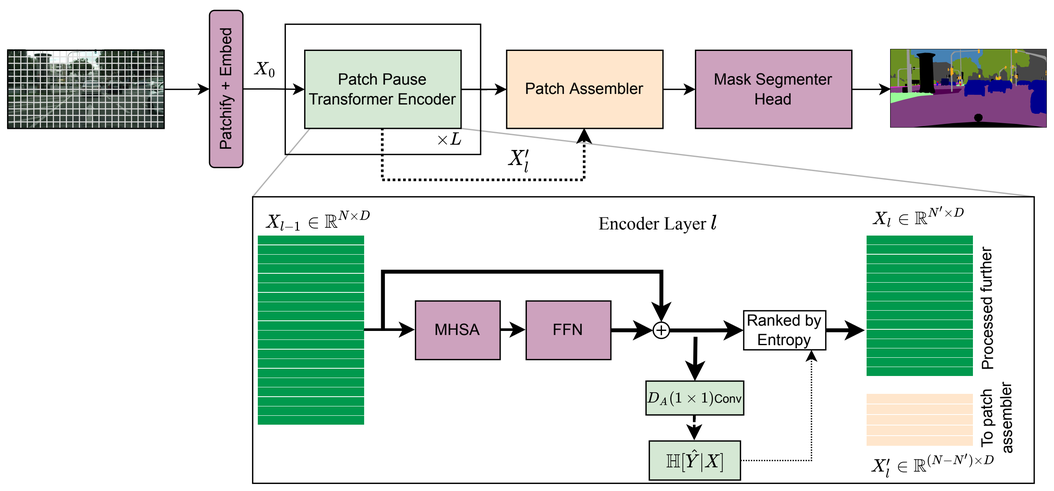


Figure: We modify Segmenter[1] to enable pausing of patches, and feed them directly to the decoder. We add a simple auxiliary 1×1 convolution decoder, and use the predicted posterior entropy $\mathbb{H}[\hat{Y}|X]$ of each component of X to reorder the feature representation X. A portion τ of this feature representation is paused and fed to the decoder directly. The rest of the features (of size $N' < N$) are processed further.

Posterior entropy correlates with correctness

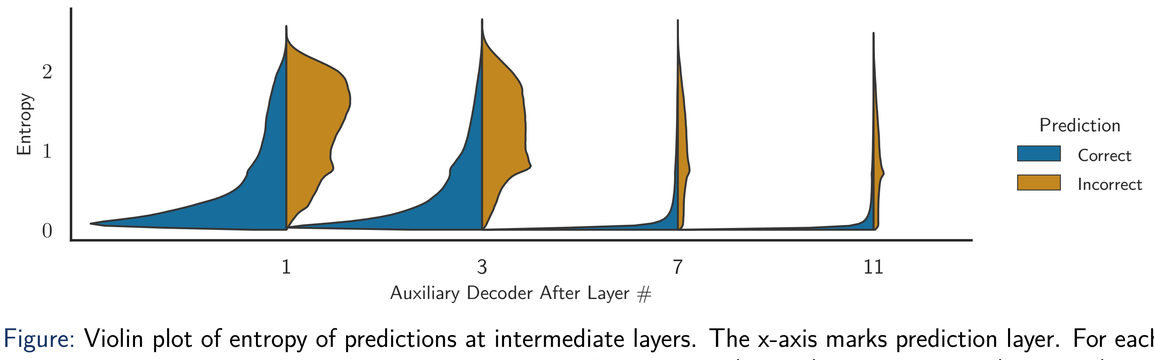


Figure: Violin plot of entropy of predictions at intermediate layers. The x-axis marks prediction layer. For each layer, the entropy distribution is shown for tokens classified correctly (in blue) and incorrectly (in orange). We see that the entropies of the predictions for tokens correctly classified accumulate in the low values in the later layers (blue spike on the bottom-right).

Acknowledgements

Research supported by the "Swiss Center for Drones and Robotics - SCDR" of the Department of Defence, Civil Protection and Sport via armasuisse S+T under project n°050-38.

One training to pause them all!: Training to handle various run-time requirements

- For each batch, sample a layer $l \sim \mathcal{U}\{L^{\text{lo}}, L^{\text{hi}}\}$ and a patch pausing proportion $\tau_l \sim \mathcal{U}[\tau_l^{\text{lo}}, \tau_l^{\text{hi}}]$.
- We employ an auxiliary decoder after the operations of layer l, and pause τ_l patches with the lowest entropy.
- Training both main and aux decoder with cross entropy.

Pausing patches at multiple layers

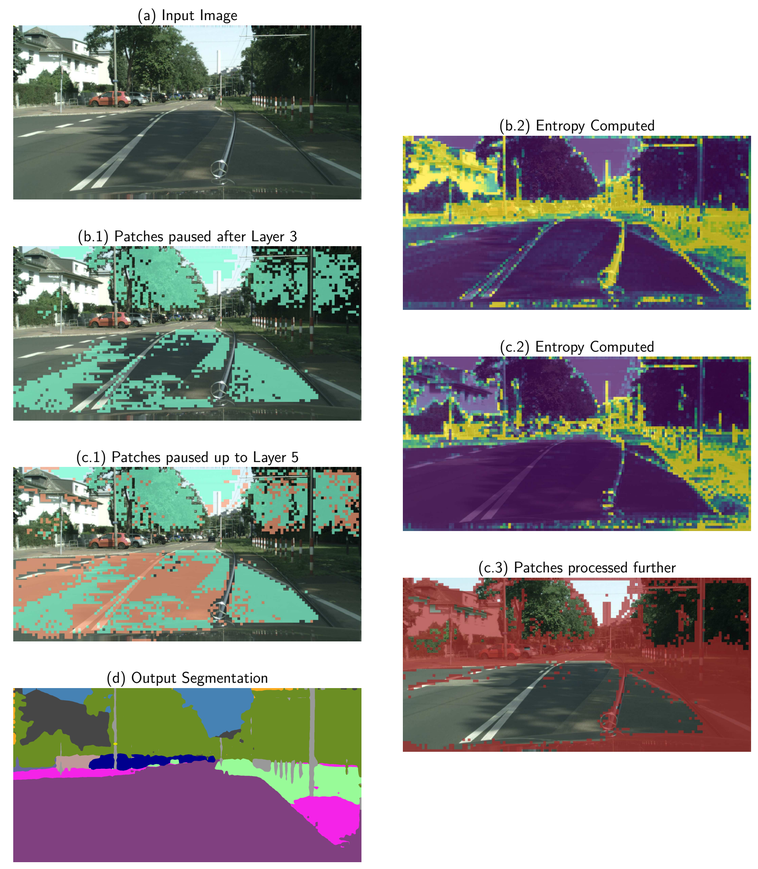


Figure: Our method progressively stops processing patches after they reach a low enough prediction entropy. (a) is the input image. (b.1, c.1) show the patches that are stopped from being processed after the 3rd and 5th layers. (b.2) and (c.2) show entropy computed from the auxiliary decoders that is used to pause patches. The network automatically pauses easy parts of the image while allocating more computation to the parts that correspond to boundaries, and to smaller and rarer classes in (c.3).

Traversing the pareto front of throughput-mIoU tradeoff

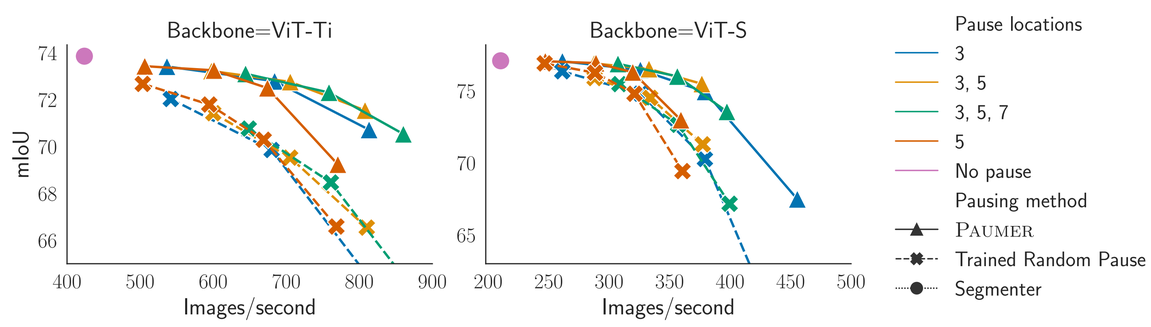


Figure: Comparing PAUMER to baseline and random pausing on Cityscapes validation set.

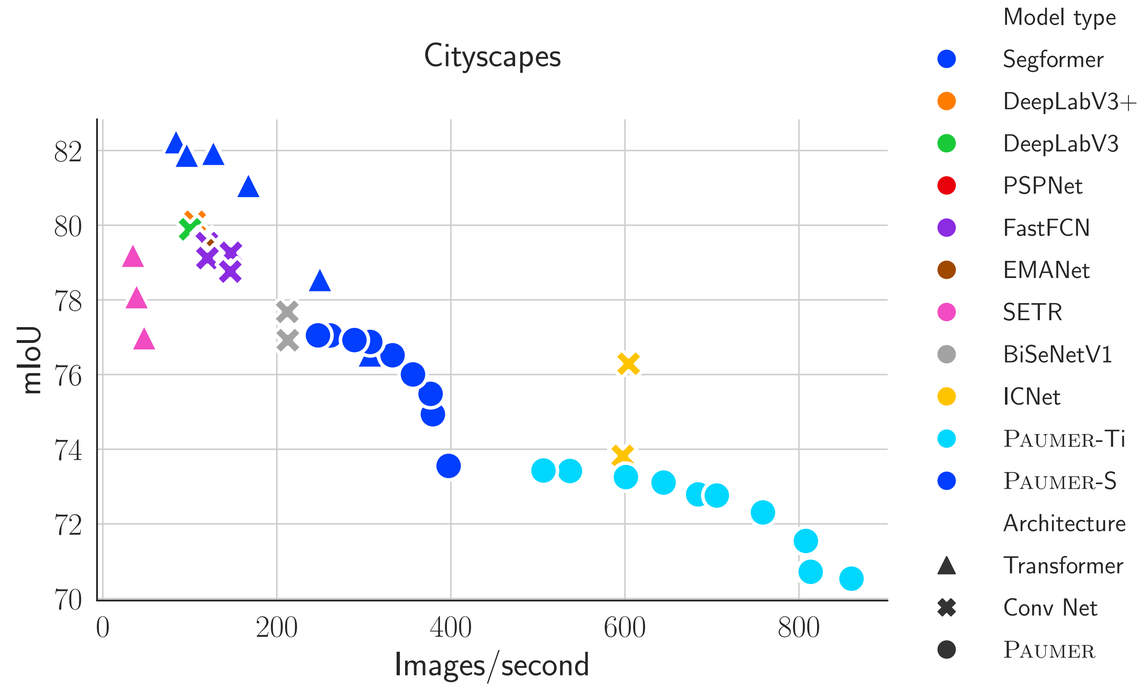


Figure: Comparing PAUMER's speed-accuracy tradeoff with other architectures.

[1] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021.