AutoFocusFormer:

Image Segmentation off the Grid

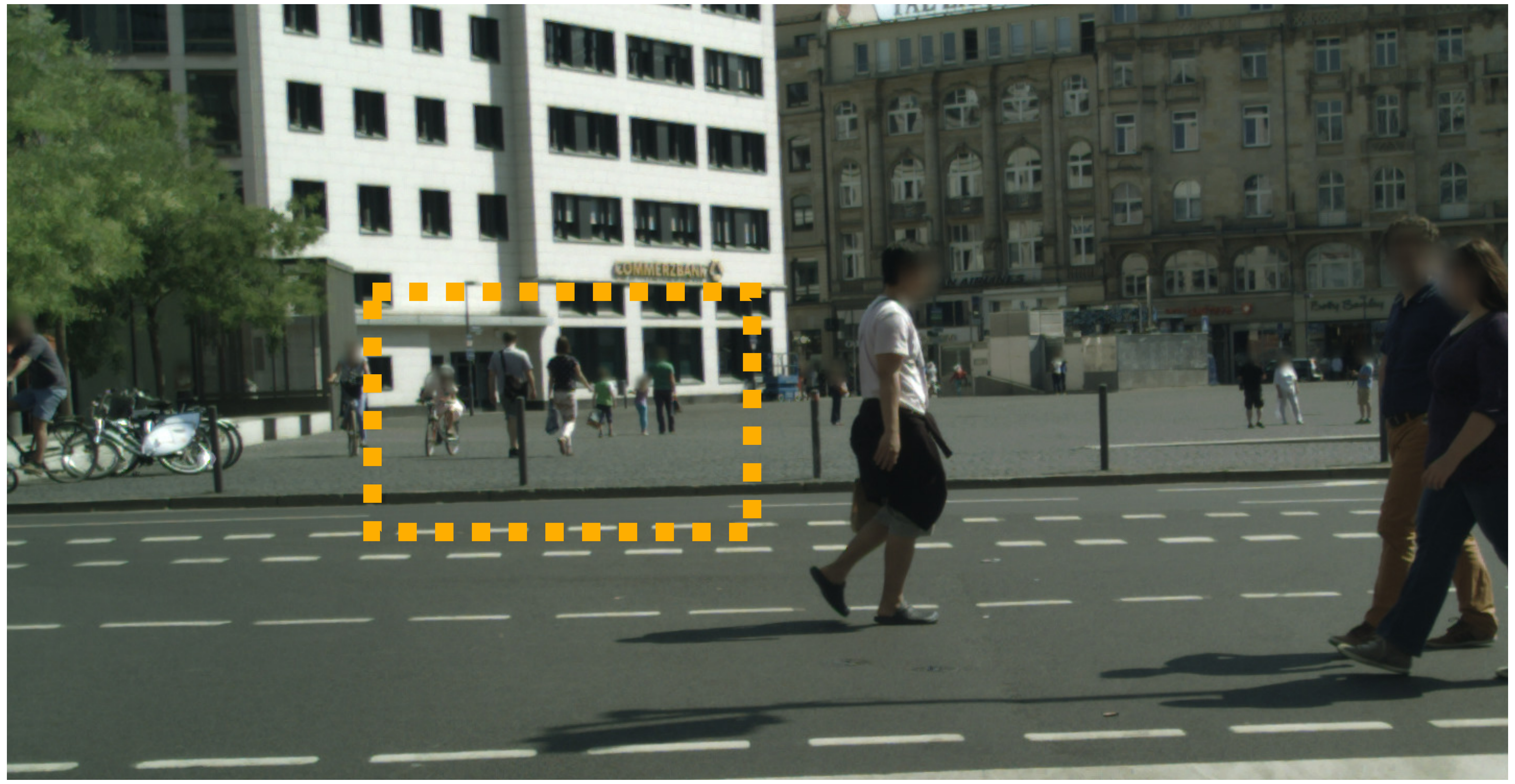
THU-AM-167

Chen Ziwen, Kaushik Patnaik, Shuangfei Zhai, Alvin Wan, Zhile Ren, Alex Schwing, Alex Colburn, Li Fuxin

2023.6 | Apple Inc. & Oregon State University

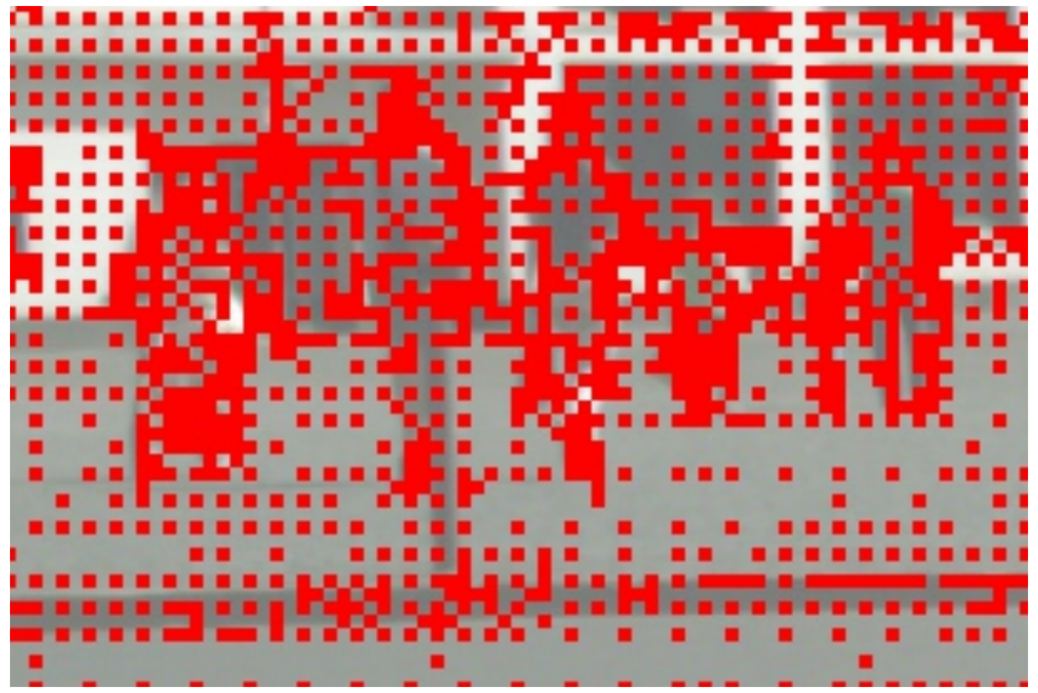
AutoFocusFormer(AFF)

Multi-stage, local-attention transformer, equipped with successive adaptive downsampling layers

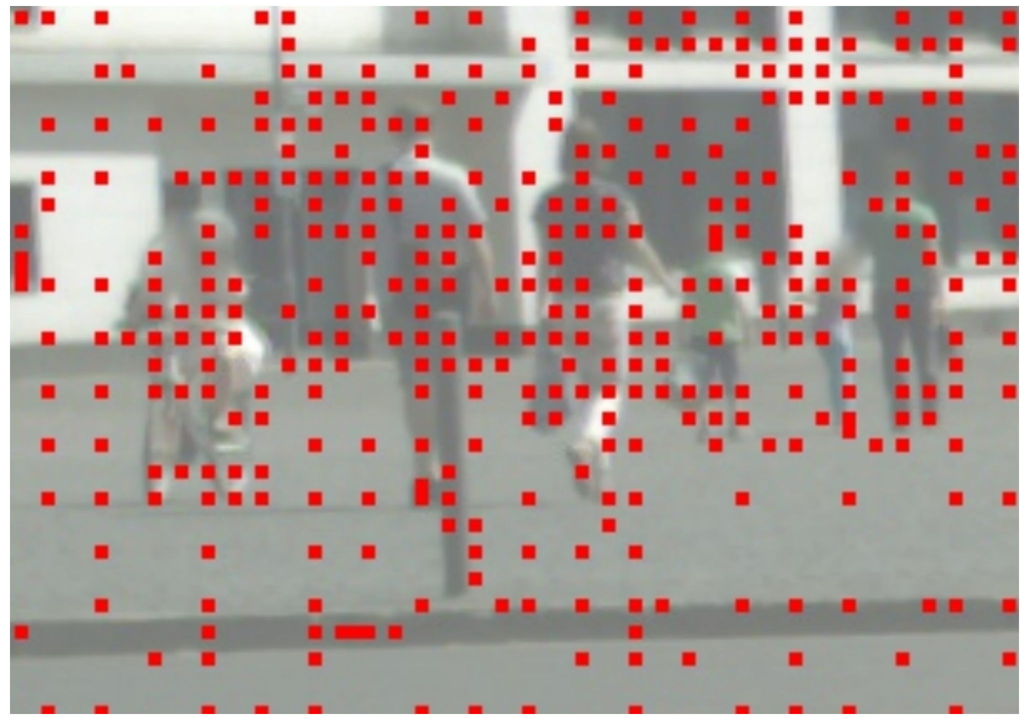


Image

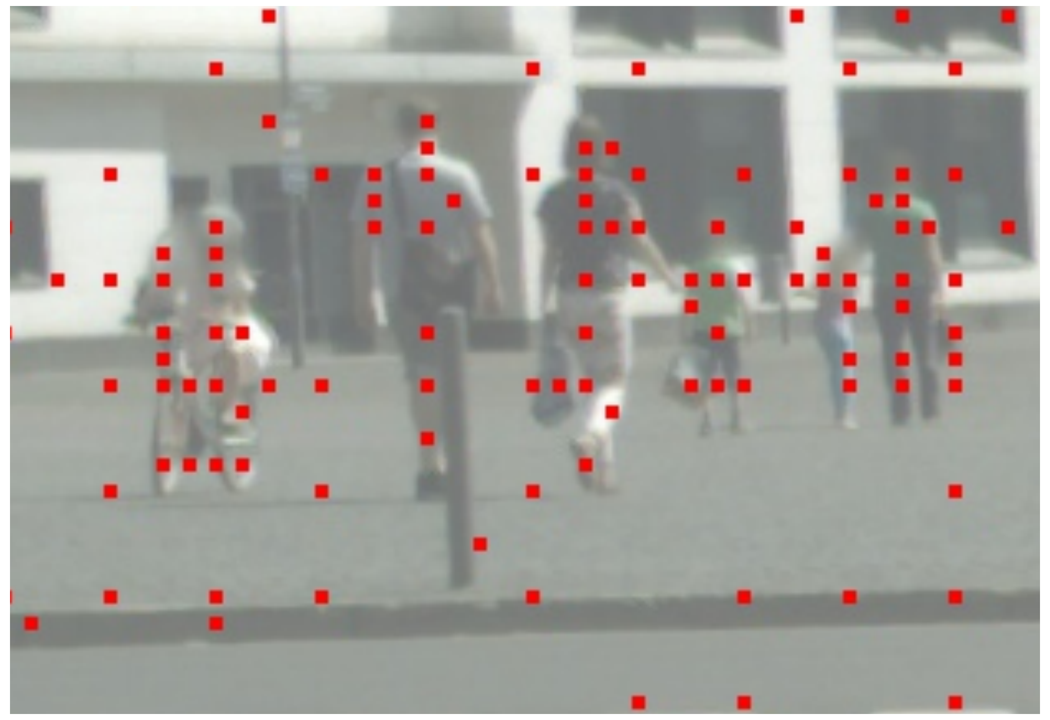
Stage 2



Stage 3

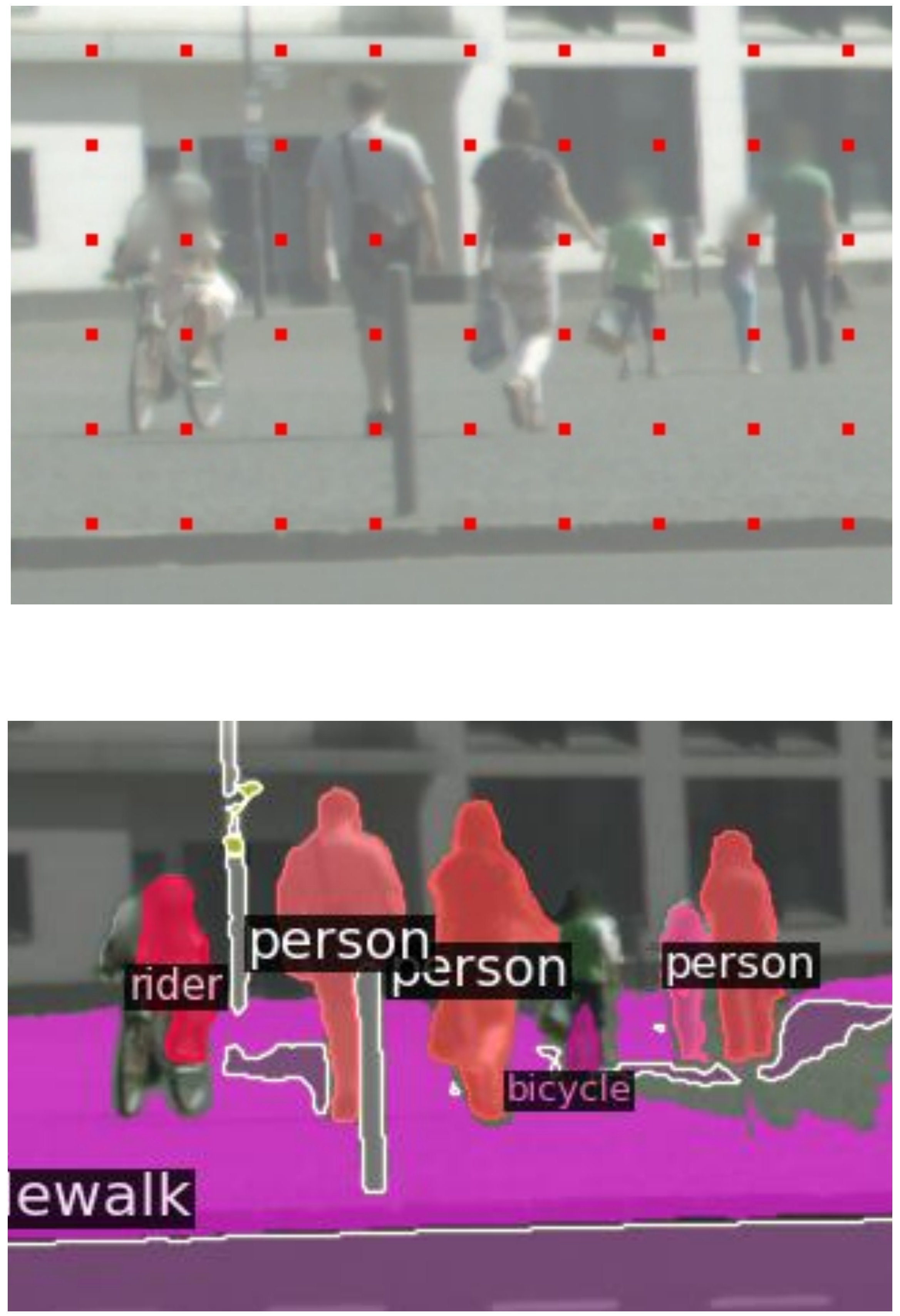


Stage 4

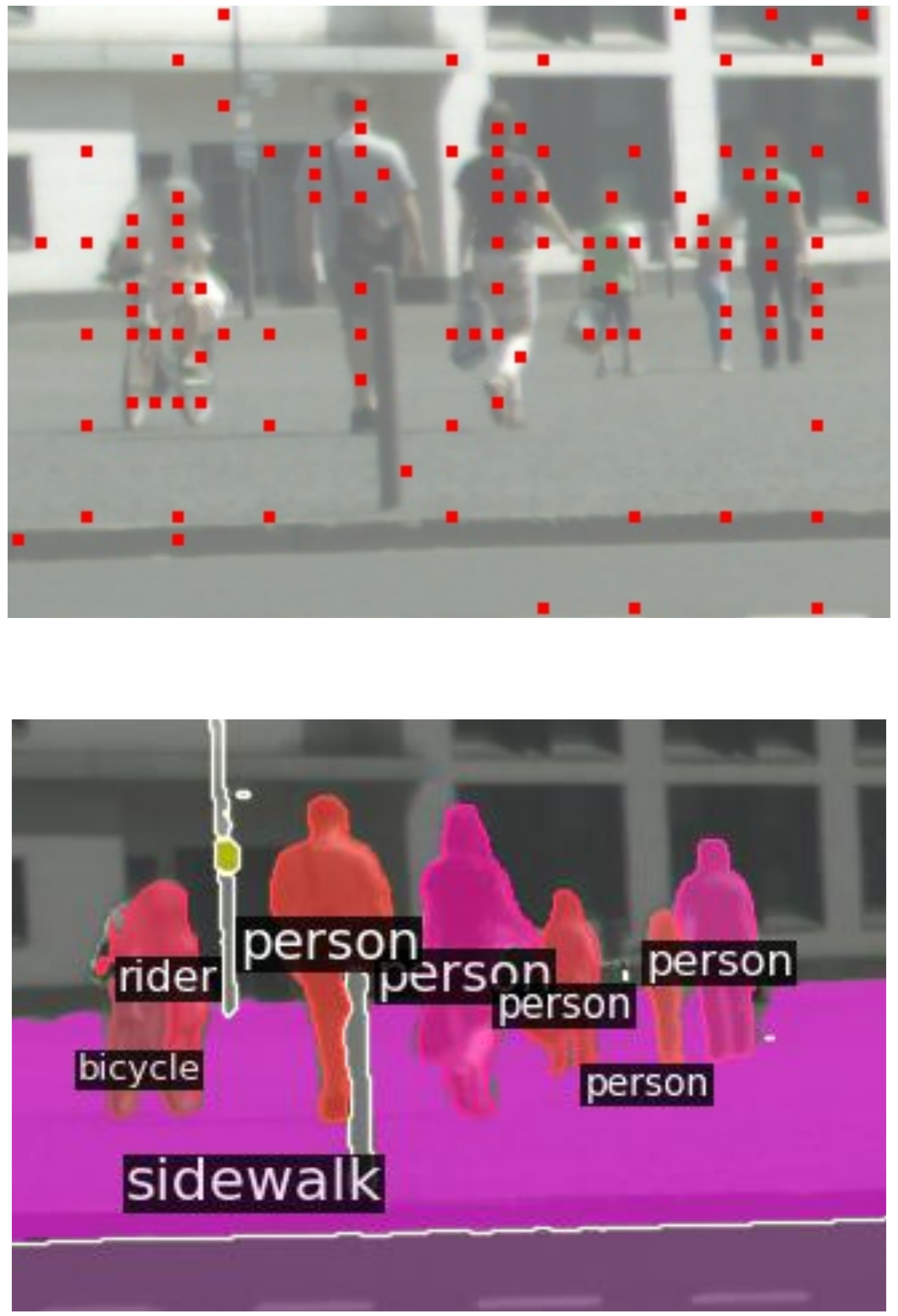


Red pixels indicate remaining tokens

AutoFocusFormer(AFF)



Panoptic Segmentation
With Swin backbone



Panoptic Segmentation
With AFF backbone

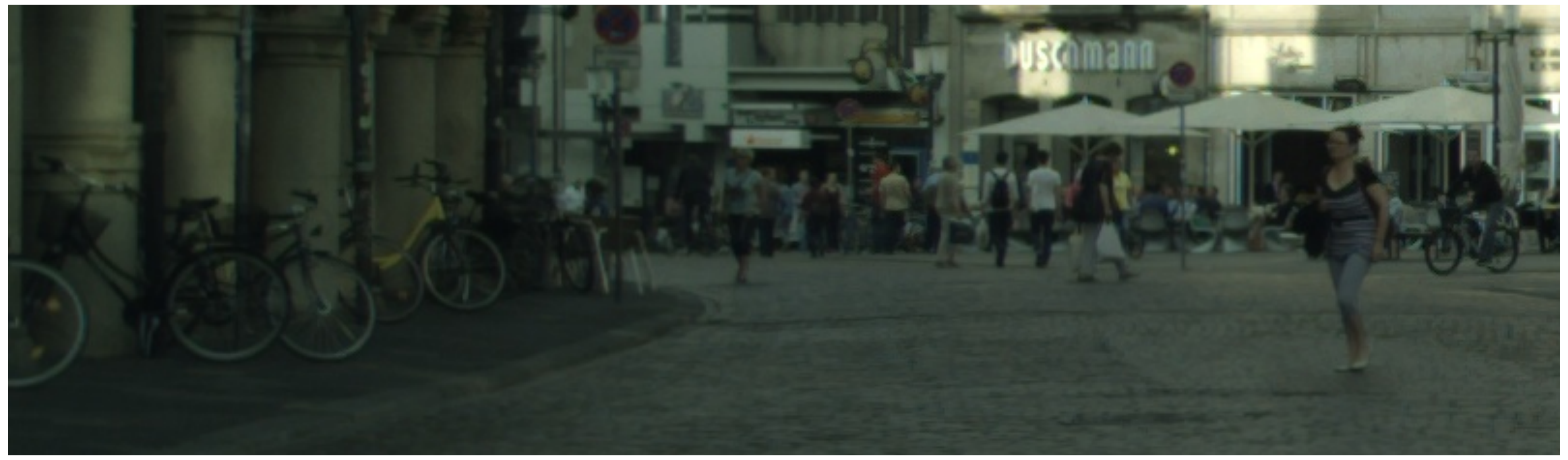
Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Result Highlights

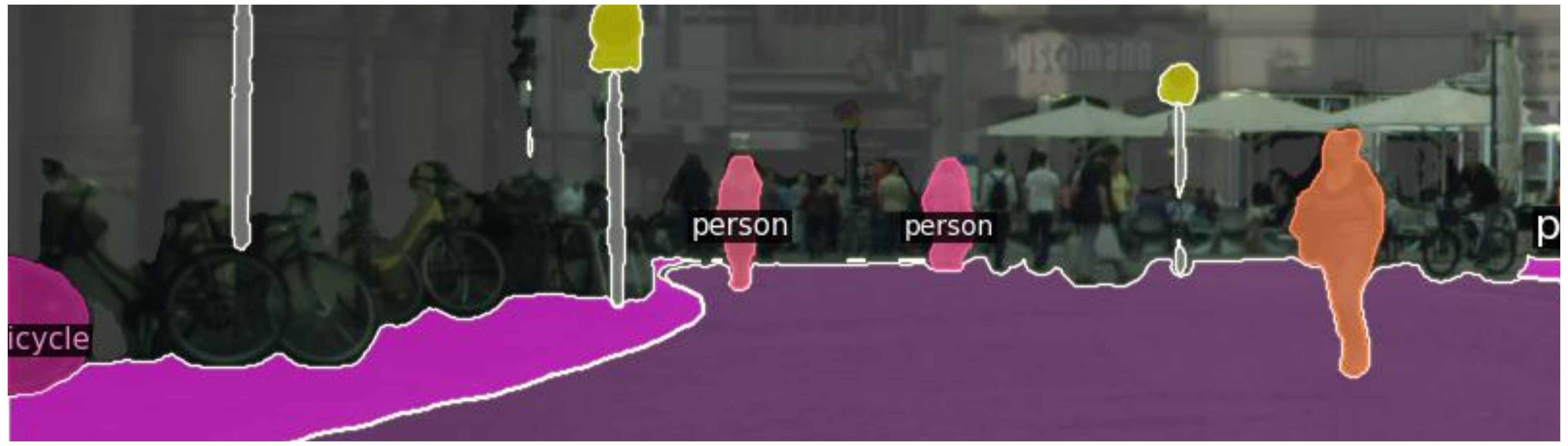
On Cityscapes segmentation

- AFF-Tiny performs on par with Swin-Base (3.3x larger)
- AFF-Small performs on par with Swin-Large (4.6x larger)

For all datasets, AFF saves ~30% FLOPs without much drop in performance



Image



SWIN prediction

AFF prediction

Motivation

Natural images often have highly imbalanced content density

Mainstream networks use grid downsampling

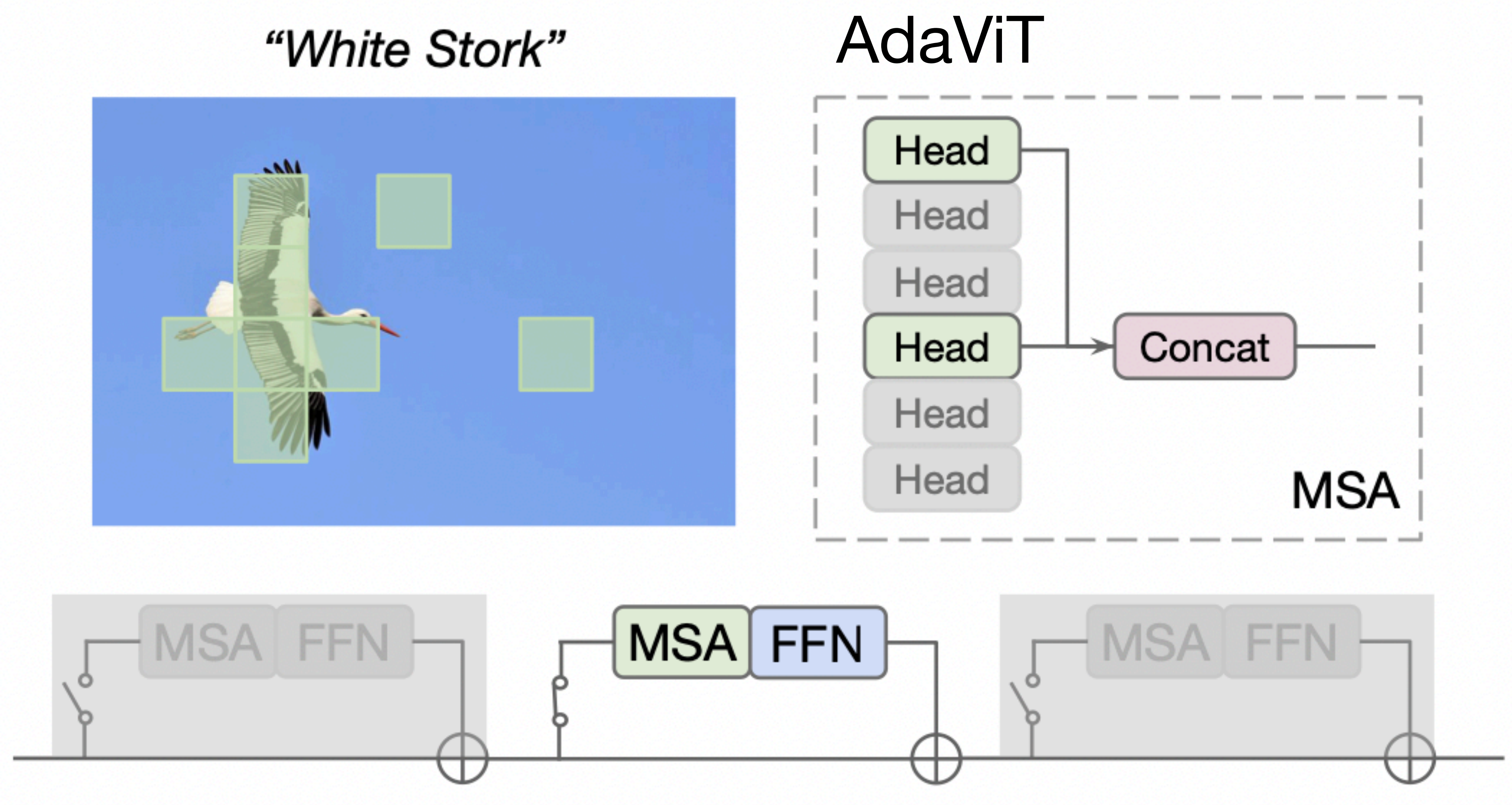
- Mis-classify small objects
- Waste computation on large objects

Adaptive Models on Transformers

E.g., AdaViT, DynamicViT, A-ViT…

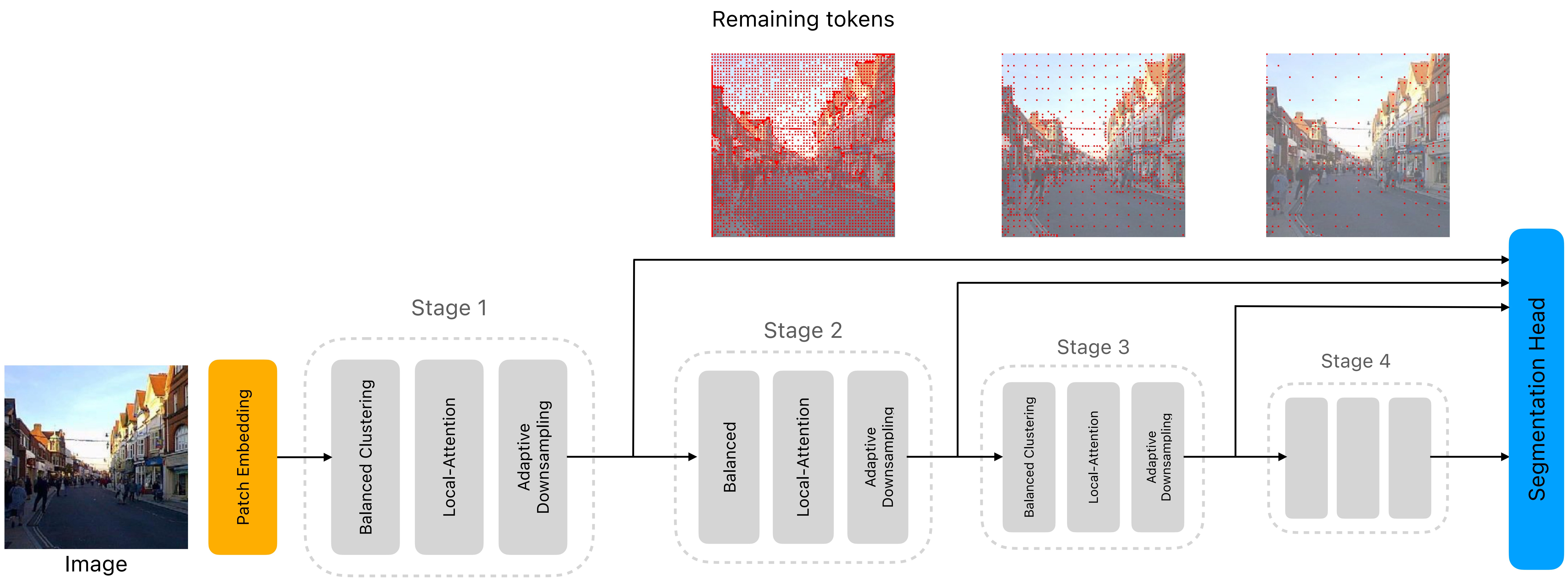
- Adopt global attention (quadratic complexity!)
- No actual downsampling in training (need gradient)

Thus, they cannot scale to high-resolution segmentation tasks!

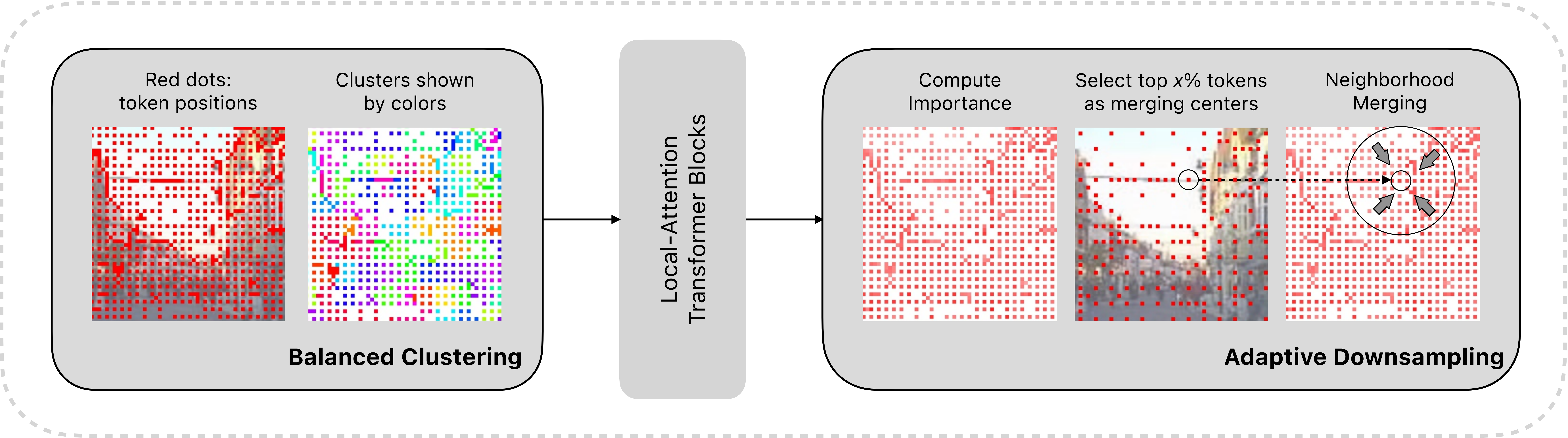


Meng, Lingchen, et al. "Adavit: Adaptive vision transformers for efficient image recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

AFF Architecture



AFF Architecture



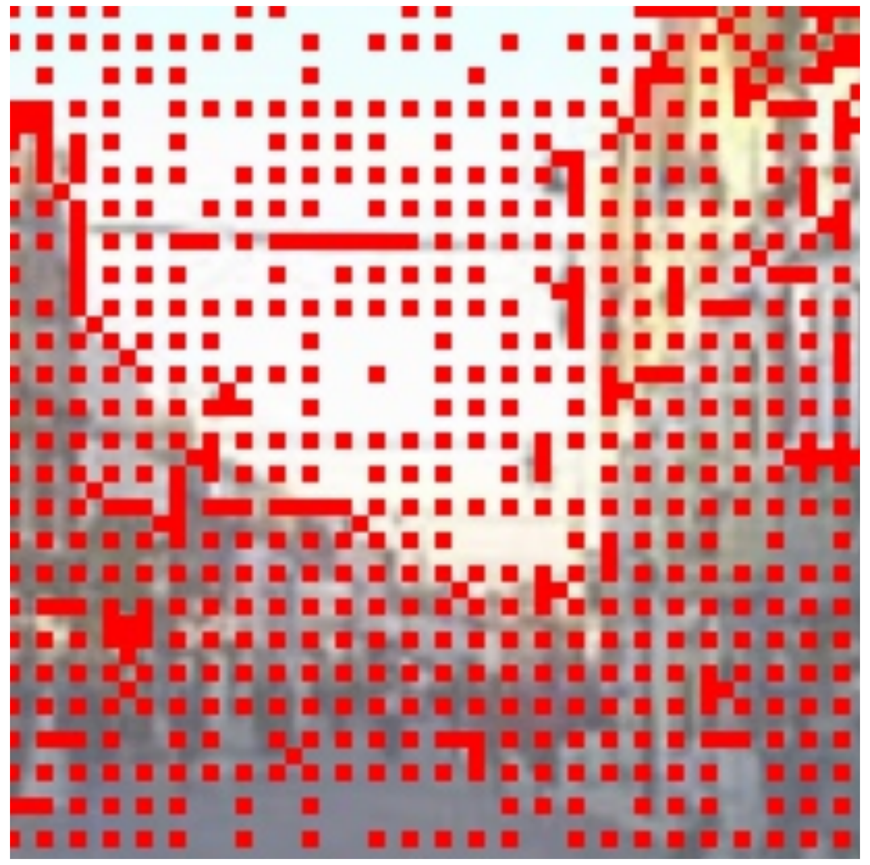
Three modules that form AFF's stage

AFF Clustering

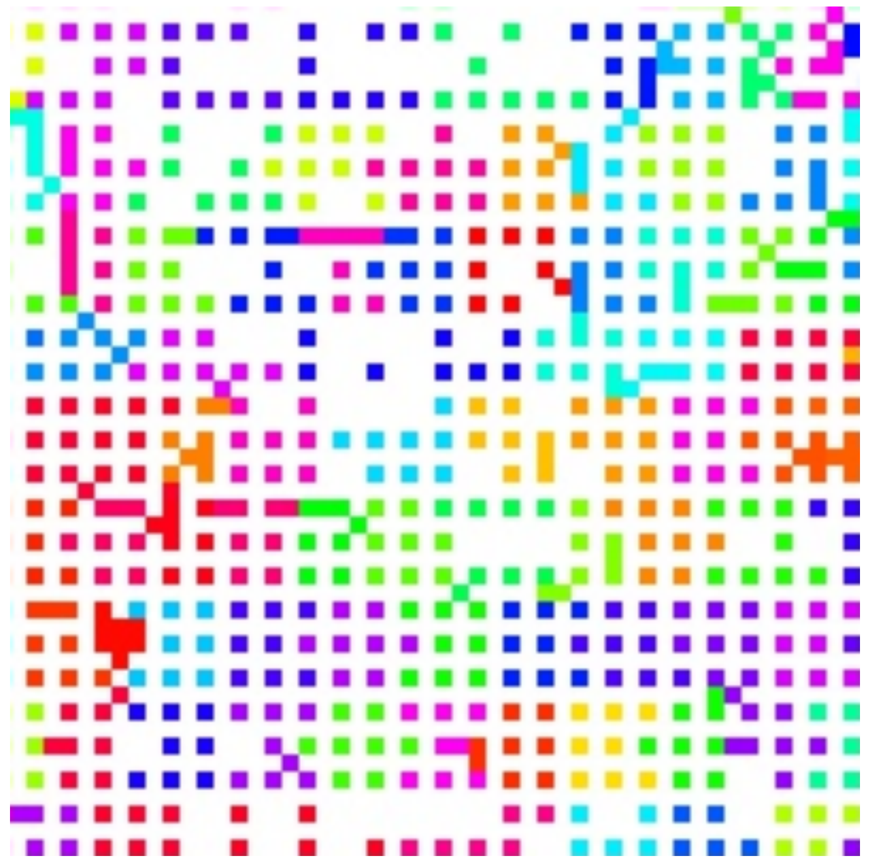
We propose a fast, non-iterative, equal-sized clustering method for 2D tokens

The method is based on space-filling curves

Please refer to our paper for algorithm details!

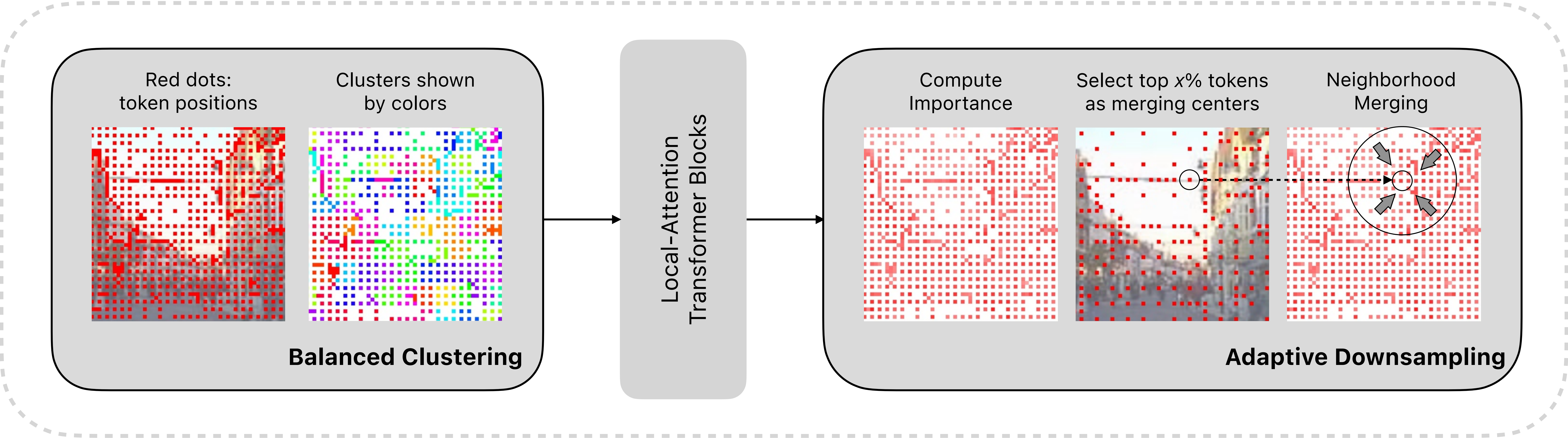


Red dots: tokens



Clusters

AFF Architecture



Three modules that form AFF's stage

AFF Local Attention

Using the equal-sized clusters, we define the neighborhood of a token by its nearest *R* clusters

Overlapping neighborhoods enable the smooth propagation of information among tokens

On-grid

Off-grid

Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

AFF Local Attention

On each neighborhood, we apply the standard transformer self-attention:

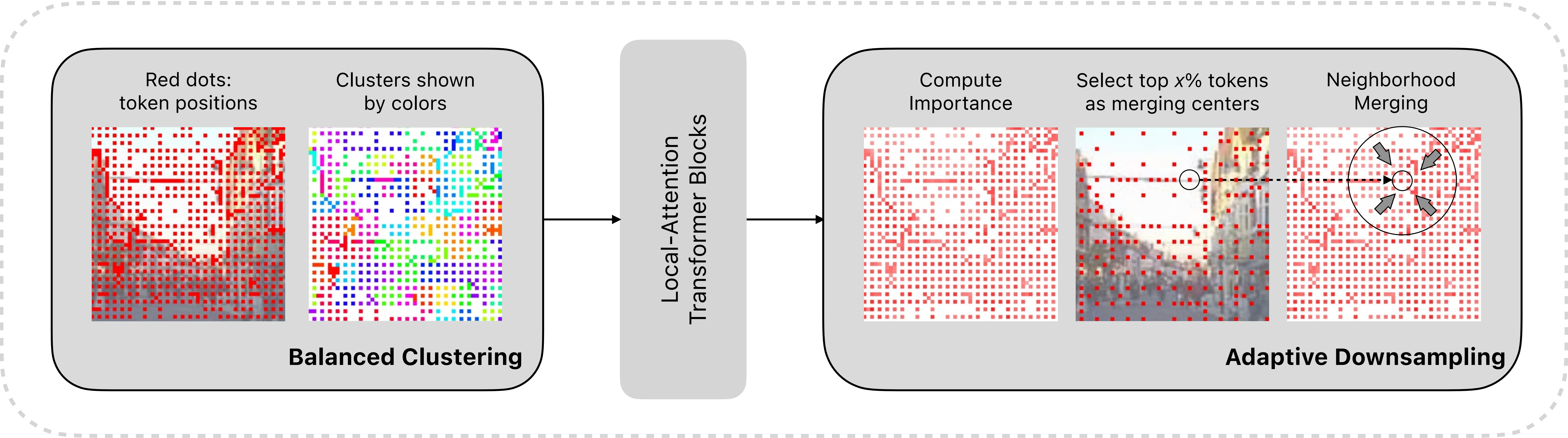
$$A = \mathrm{softmax}(QK^T + P)$$

where $P_{i,j} = w(p_i - p_j)$ is the **positional embedding**

We further make this embedding aware of potential **rotation/scale invariances** by expanding the relative position vector:

$$\left(\Delta x, \Delta y, \sqrt{\Delta x^2 + \Delta y^2}, \frac{\Delta x}{\sqrt{\Delta x^2 + \Delta y^2}}, \frac{\Delta y}{\sqrt{\Delta x^2 + \Delta y^2}}\right)$$

AFF Architecture



Three modules that form AFF's stage

Adaptive Downsampling

First, for each token, we compute an importance score s_i

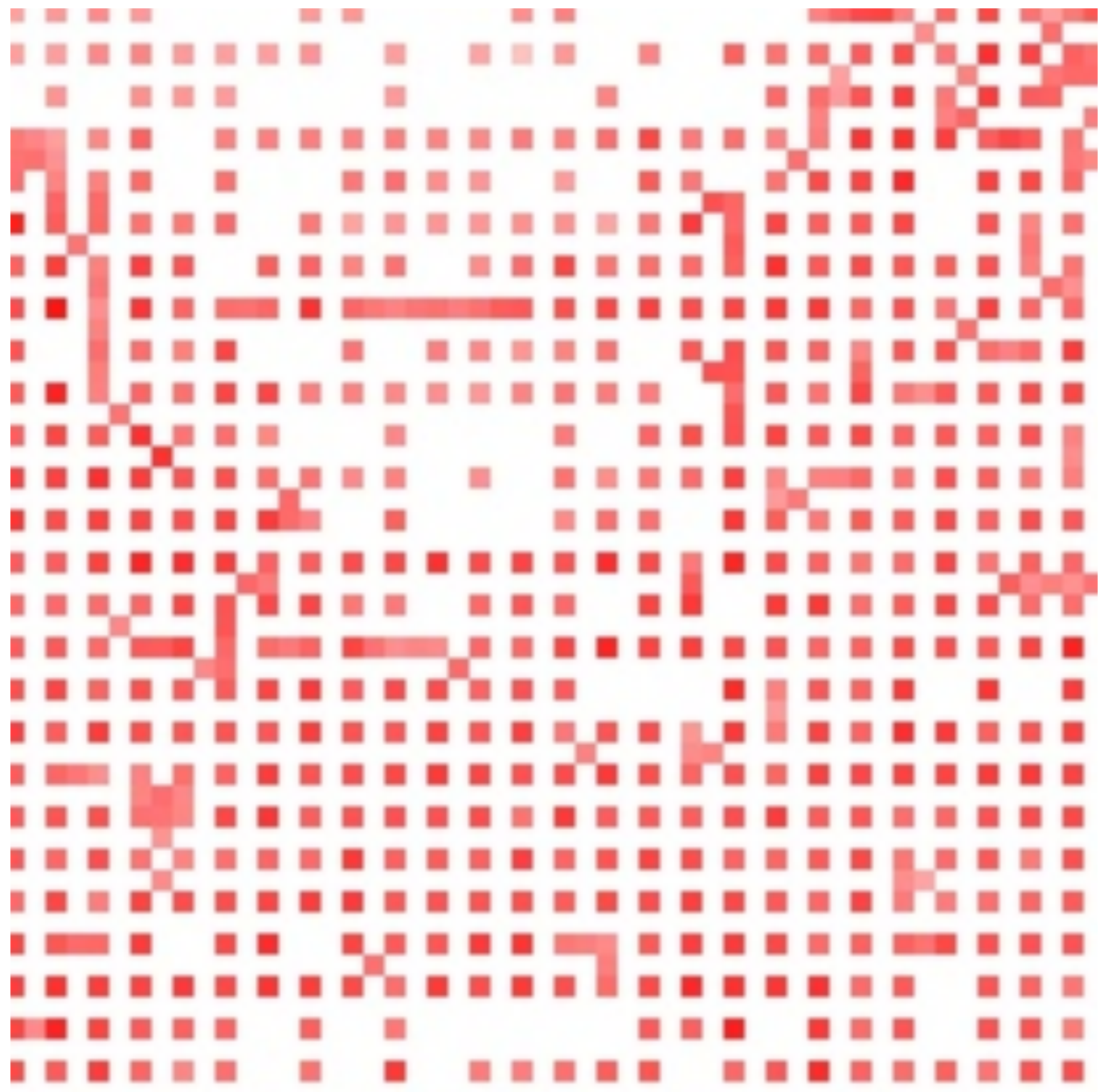
f_i: token feature

l: a fully-connected layer

Importance score = $\overbrace{\sigma(l(f_i))}^{s_i}$.

Sigmoid

Compute Importance

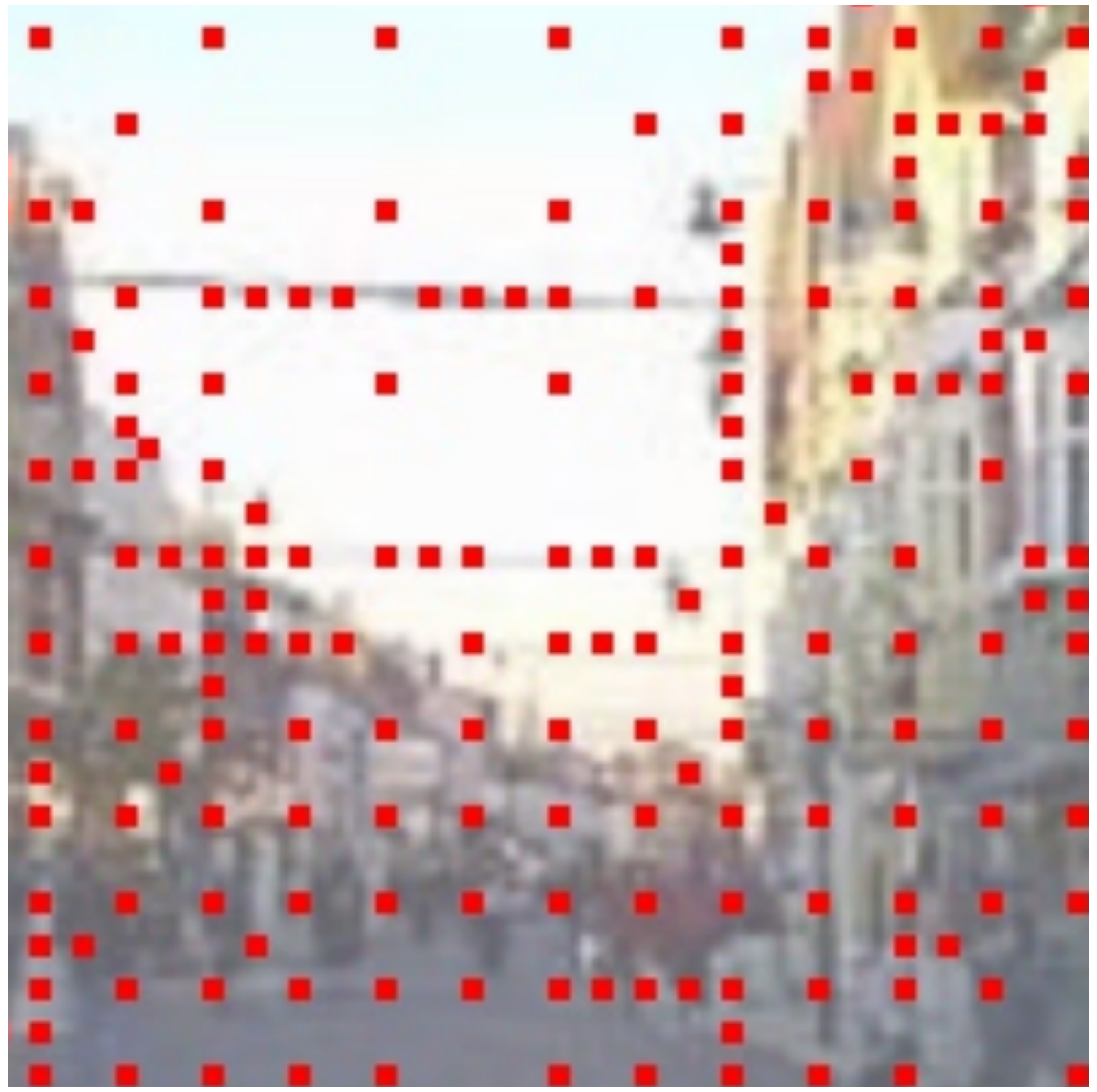


Merging center selection

Second, we select top *x*% tokens according to the importance scores

x% is the downsampling rate (we show experiment results with 1/4 and 1/5)

Select top *x*% tokens

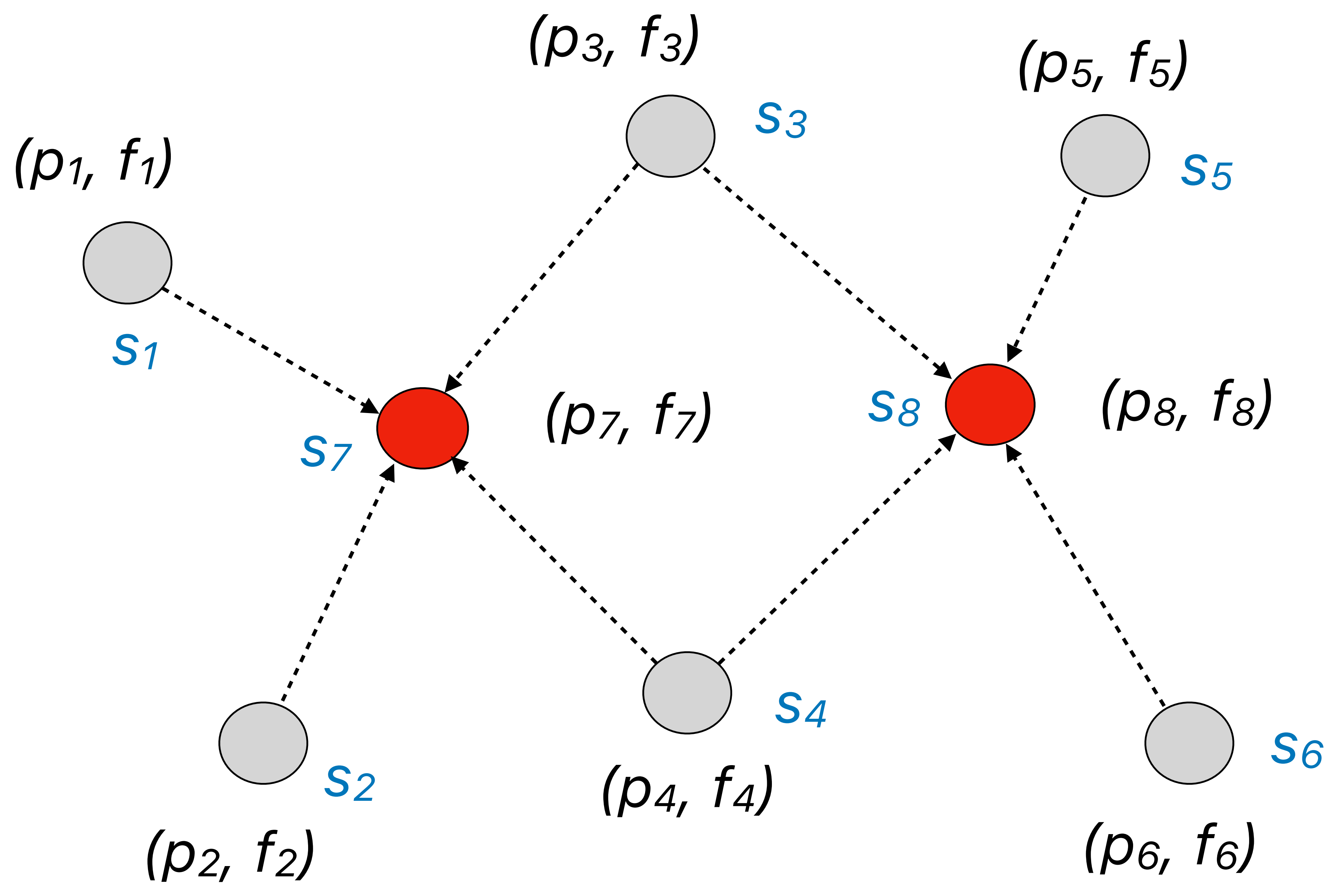


Neighborhood Merging

Lastly, we merge the neighborhoods of the selected tokens

We use a PointConv layer, modulated by the learnable importance score s_i's

The output is x% merged tokens



Wu, Wenxuan, Zhongang Qi, and Li Fuxin. "Pointconv: Deep convolutional networks on 3d point clouds." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2019.

Experiments

ImageNet classification

1/5: downsampling rate

Model	Top-1 Acc	# Params	FLOPs
Swin-Mini	76.9%	6.76M	1.07G
AFF-Mini	**78.2%**	6.75M	1.08G
AFF-Mini-1/5	77.5%	6.75M	**0.72G**
Swin-Tiny	81.3%	28M	4.5G
AFF-Tiny	**83%**	27M	4G
AFF-Tiny-1/5	82.4%	27M	**2.74G**
Swin-Small	83%	50M	8.7G
AFF-Small	**83.5%**	42.6M	8.16G
AFF-Small-1/5	**83.4%**	42.6M	**5.69G**

Experiments

ADE20K semantic segmentation

Model	mIoU	FLOPs
Swin-Mini	44.1	48.9G
AFF-Mini	**46.5**	48.3G
AFF-Mini-1/5	46.0	**39.9G**
Swin-Tiny	47.7	74G
AFF-Tiny	**50.2**	64.6G
AFF-Tiny-1/5	**50.0**	**51.1G**
Swin-Small	51.3	98G
AFF-Small	51.2	87G
AFF-Small-1/5	**51.9**	**67.2G**

Segmentation head: Mask2Former

Experiments

Cityscapes instance & panoptic segmentation

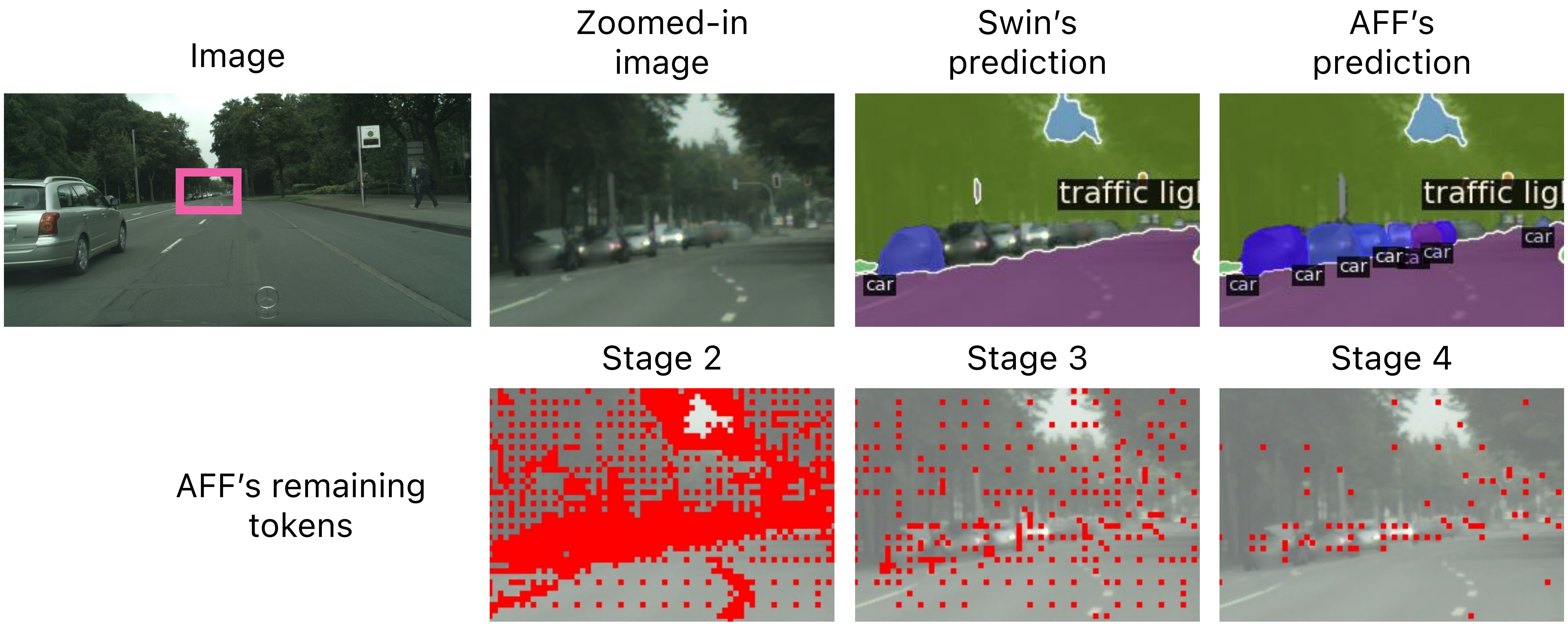
Model	Instance AP	Panoptic PQ (s.s.)	Backbone # Params
AFF-Mini	40.0	62.7	6.75M
Swin-Tiny	39.7	63.9	28M
AFF-Tiny	**42.7**	**65.7**	27M
Swin-Small	41.8	64.8	50M
AFF-Small	**44.0**	**66.9**	42.6M
Swin-Base	42	66.1	88M
Swin-Large	43.7	66.6	197M

3.3x

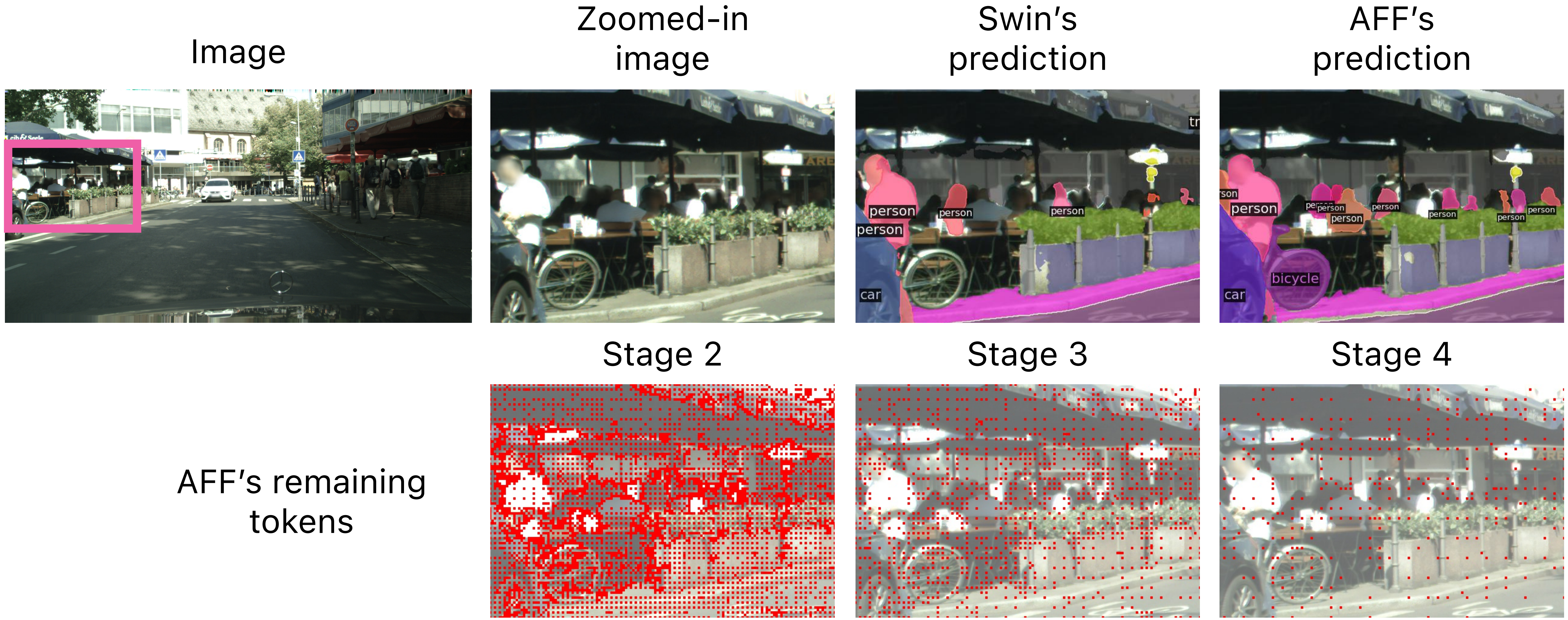
4.6x

Segmentation head: Mask2Former

Qualitative results

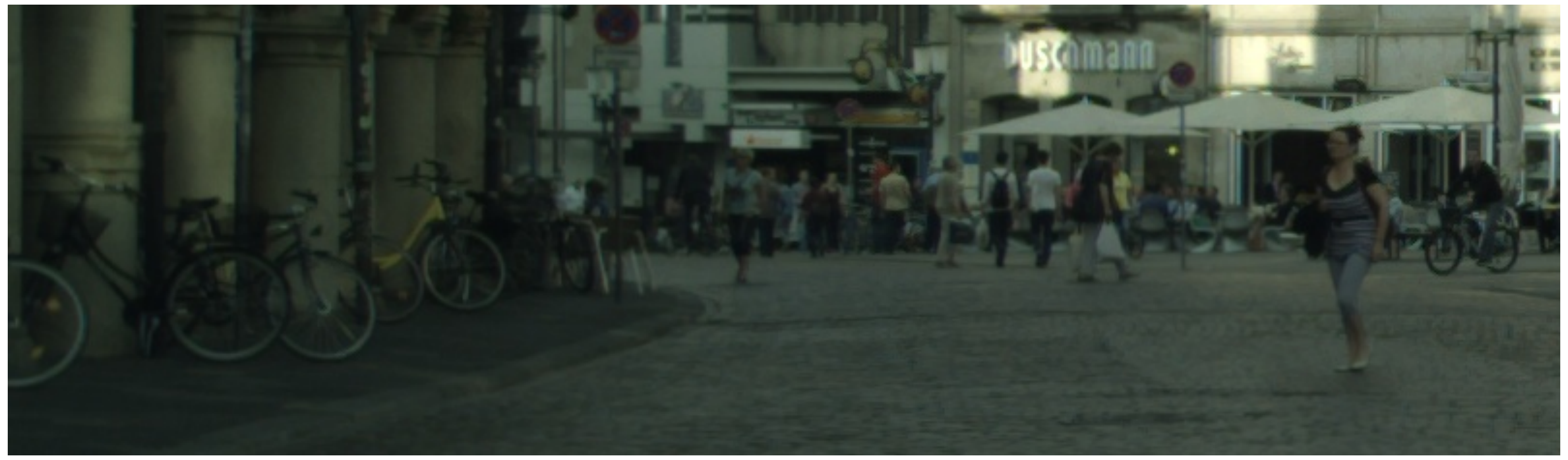


Qualitative results

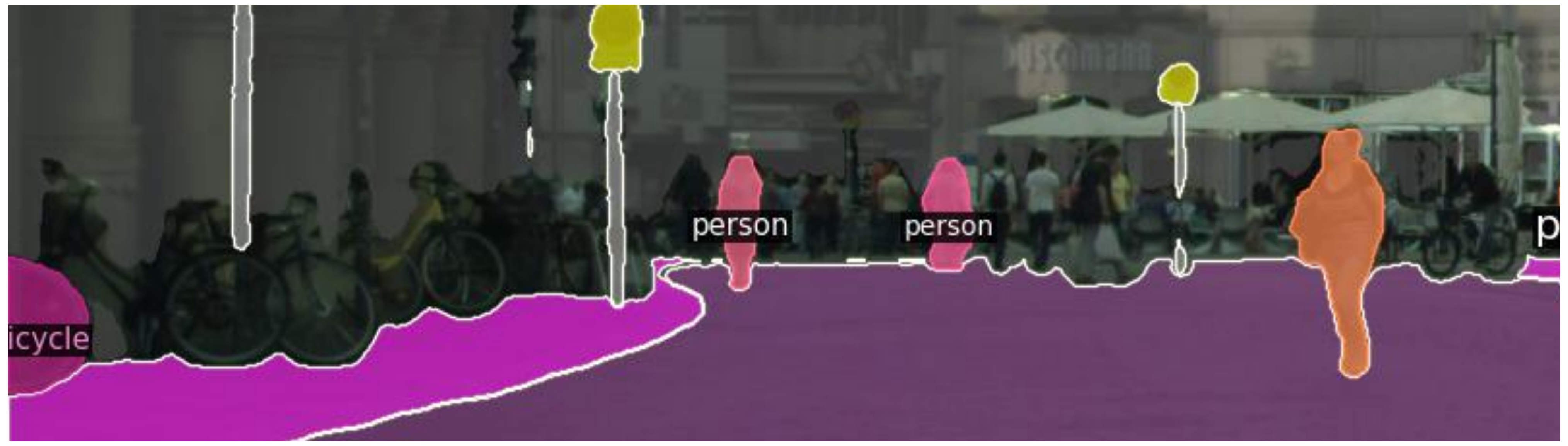


Conclusions

- We introduce the first adaptive-downsampling network capable of dense prediction tasks such as semantic/instance segmentation
- Flexible downsampling rate (e.g., 1/5 vs. traditional 1/4)
- Significant savings on FLOPs and significant improvement on recognition of small objects



Image



SWIN prediction

AFF prediction