A Spitting Image: Modular Superpixel Tokenization in Vision Transformers

M. Aasan, O. Kolbjørnsen, A. Schistad Solberg, and A. Ramírez Rivera

University of Oslo, SFI Visual Intelligence, AkerBP

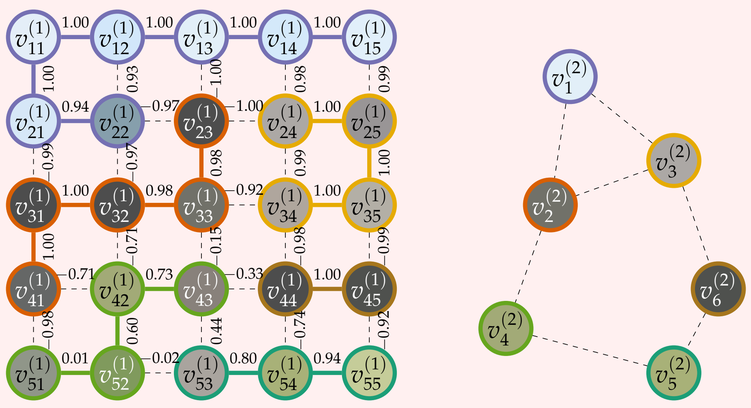
Motivation

Goal: Adaptable tokenization for ViTs

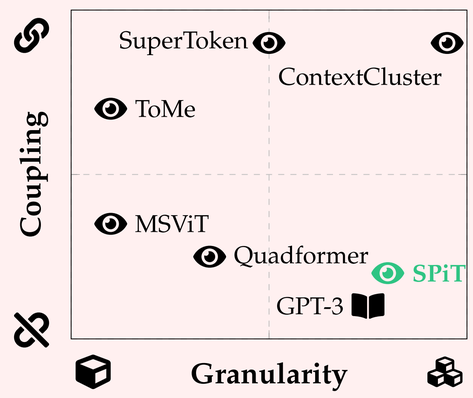
Flexible tokenization can more readily adapt to image content and context

- Superpixels generalize patch-based tokenization in ViTs
- Efficent tokenization via hierarchical graph-based superpixels
- Redined spatial resolution with pixel-level granularity for dense predictions

Superpixel Aggregation



Taxonomy



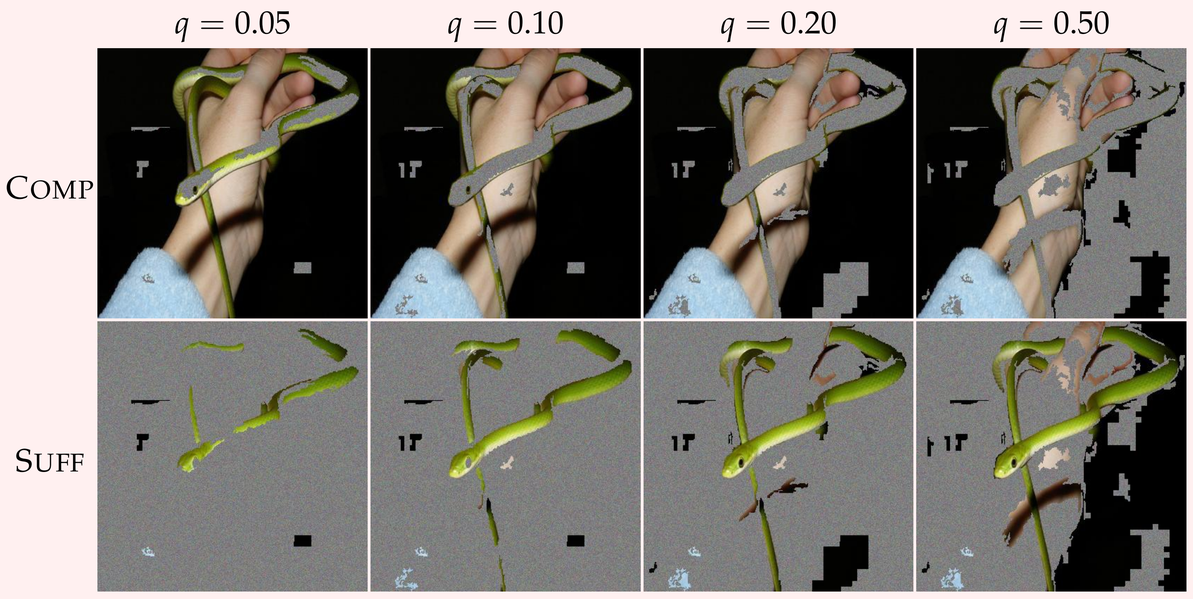
Classification Accuracy

Model			INReaL		IN1k		Caltech256		Cifar100	
Name	Grad.	Im./s.	Lin.	kNN	Lin.	kNN	Lin.	kNN	Lin.	kNN
ViT-B16	✗	793.04	0.85	0.85	0.80	0.74	0.88	0.88	0.89	0.90
ViT-B16	✓	721.12	0.85	0.84	**0.81**	0.75	**0.89**	0.89	**0.90**	**0.90**
RViT-B16	✗	619.86	0.84	0.83	0.79	0.72	0.87	0.88	0.89	0.84
RViT-B16	✓	585.64	0.84	0.84	0.79	0.73	0.86	0.86	0.89	0.76
SPiT-B16	✗	690.72	0.79	0.82	0.76	0.57	0.83	0.83	0.81	0.63
SPiT-B16	✓	640.59	**0.86**	**0.85**	0.80	**0.75**	0.89	**0.89**	0.88	0.85

Occlusion and Faithfulness

$$\textsc{Comp}_{Q|x,\Phi} = \frac{1}{|Q|}\sum_{q\in Q}\big(\Phi(x) - \Phi(x \setminus x_{>q})\big),$$

$$\textsc{Suff}_{Q|x,\Phi} = \frac{1}{|Q|}\sum_{q\in Q}\big(\Phi(x) - \Phi(x \setminus x_{\leq q})\big).$$

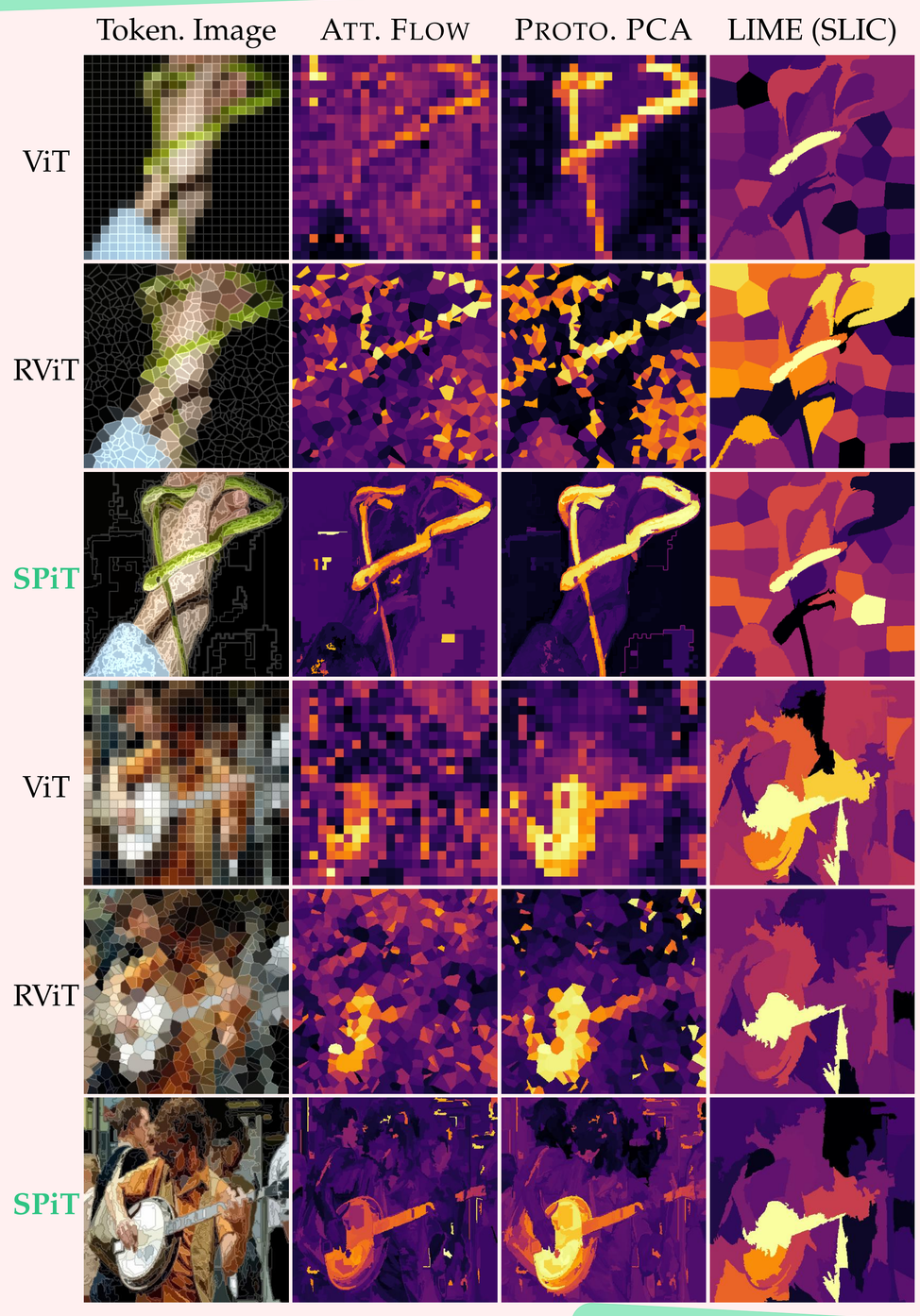


Faithfulness of Attributions

	ViT-B16		RViT-B16		SPiT-B16	
	Comp ↑	Suff ↓	Comp ↑	Suff ↓	Comp ↑	Suff ↓
LIME/SLIC	**0.244**	**0.543**	**0.236**	**0.591**	0.244	**0.520**
Att.Flow	0.160	0.664	0.223	0.685	**0.259**	0.558
Prot.PCA	0.206	0.710	0.209	0.691	0.256	0.592

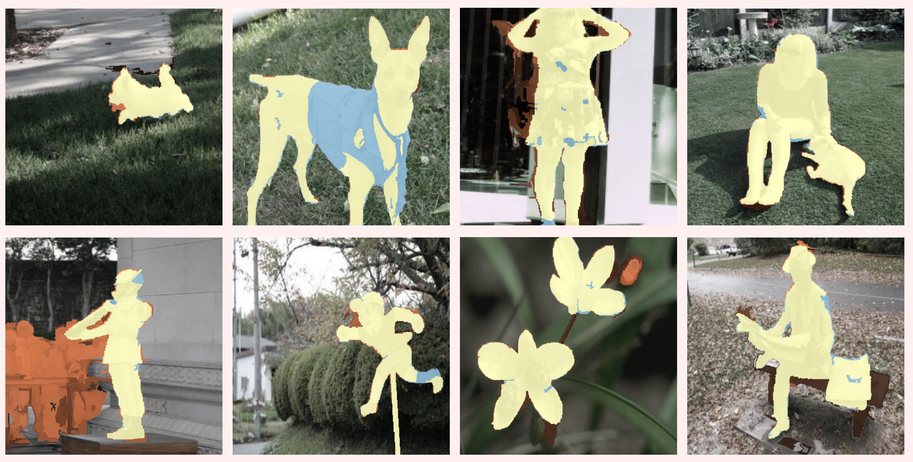
Project Page

Attributions

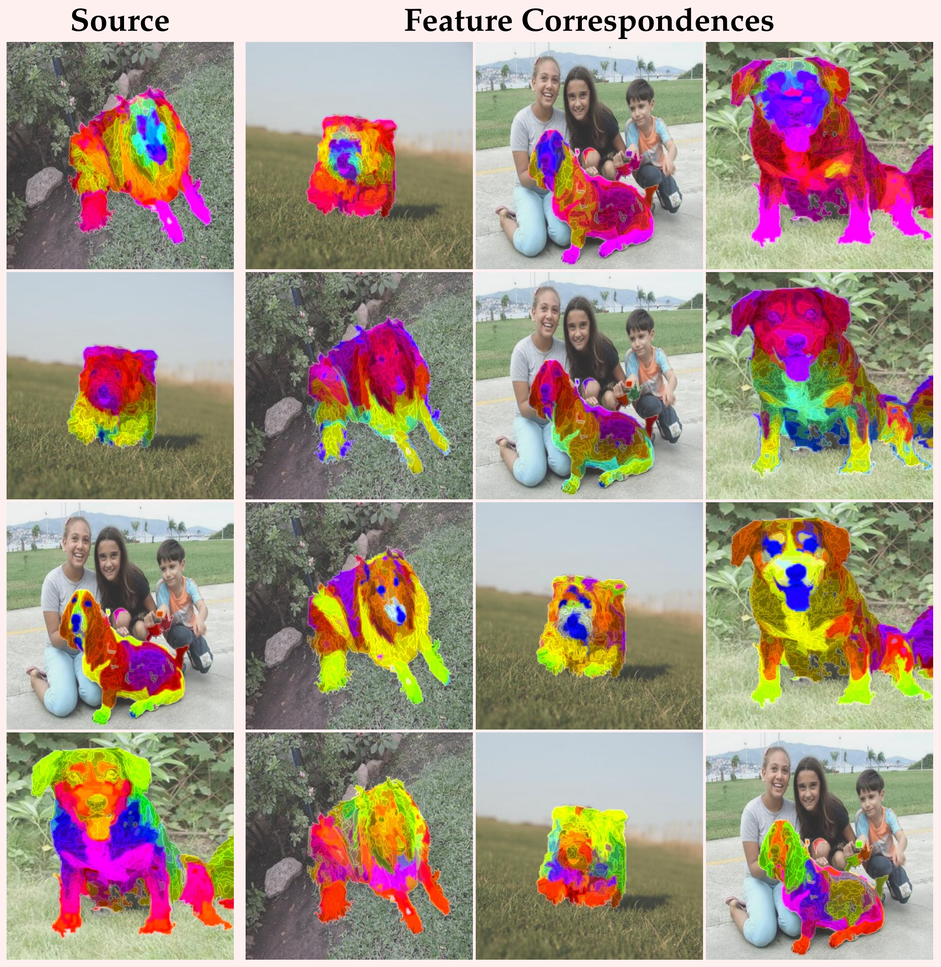


Out-of-the-box

Unsupervised Salient Segmentation



True positives | False positives | False negatives



CVF/ECCV More Exploration, Less Exploitation (MELEX) 2024 Workshop

Acknowledgements: Sigma2 NN8104K & RCN 309439