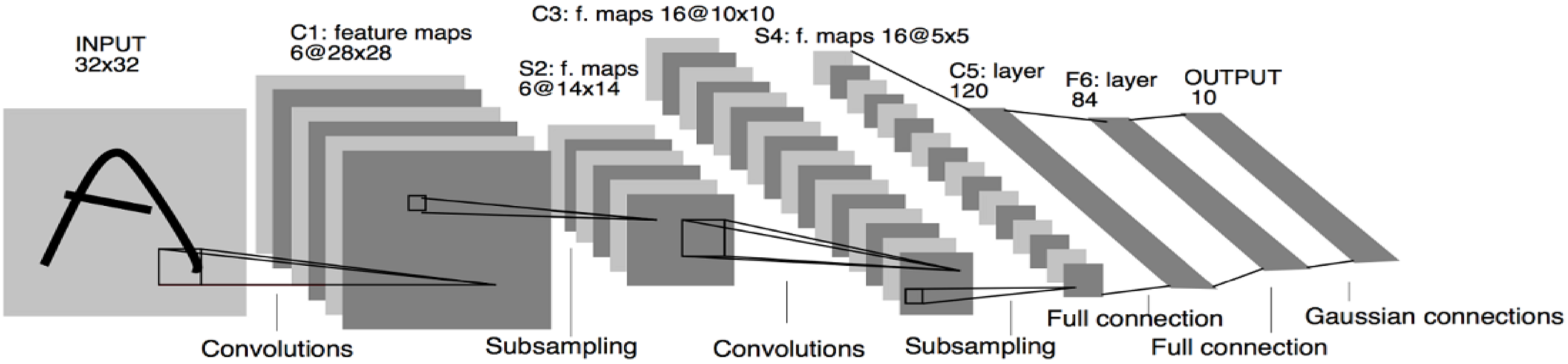
Enhancing Vision Transformer with Superpixel Representation

Pixel Representation

- Typically have high-resolution
- Need a local sliding window approach for efficient processing
- **Intractable for global self-attention**, due to the **quadratic complexity**



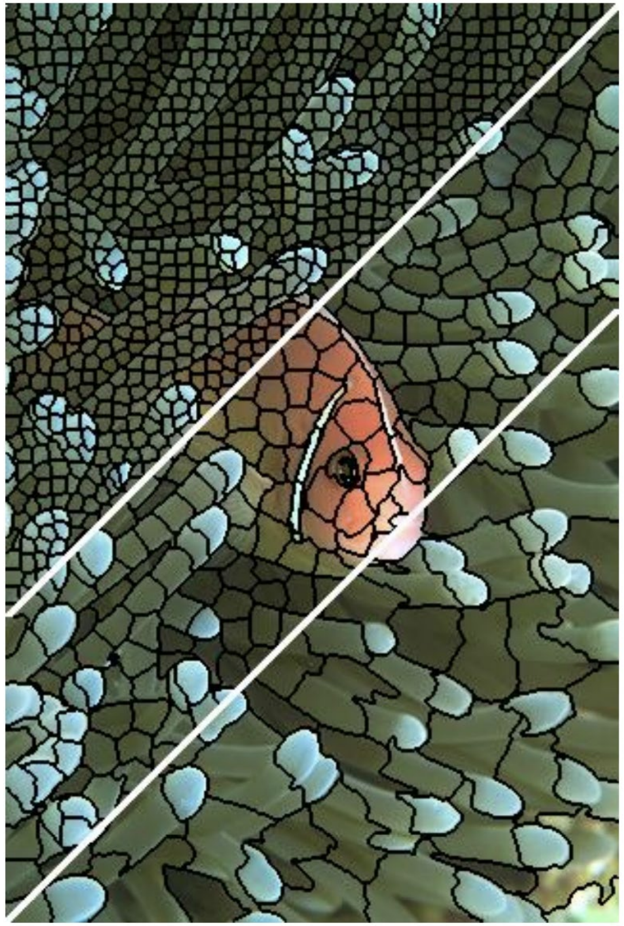
Patch Representation

- Image as a set of 16x16 patches, which enables us to learn global information
- Low-resolution, thus **sacrifices image details.**

Can superpixels help?

Superpixels **over-segment the image** into similar regions.

Usually used as the **preprocessing** step to reduce the complexity



Achanta, Radhakrishna, et al.
SLIC superpixels. 2010.

Boyan Bonevet al. Bottom-Up Processing in Complex Scenes. In Recent Progress in Brain and Cognitive Engineering, 2015.
Ming-Yu Liu et al. Entropy rate superpixel segmentation. In CVPR, 2011.

Superpixel Representation

- Superpixel features
- Superpixel association

Pixel Unfolding

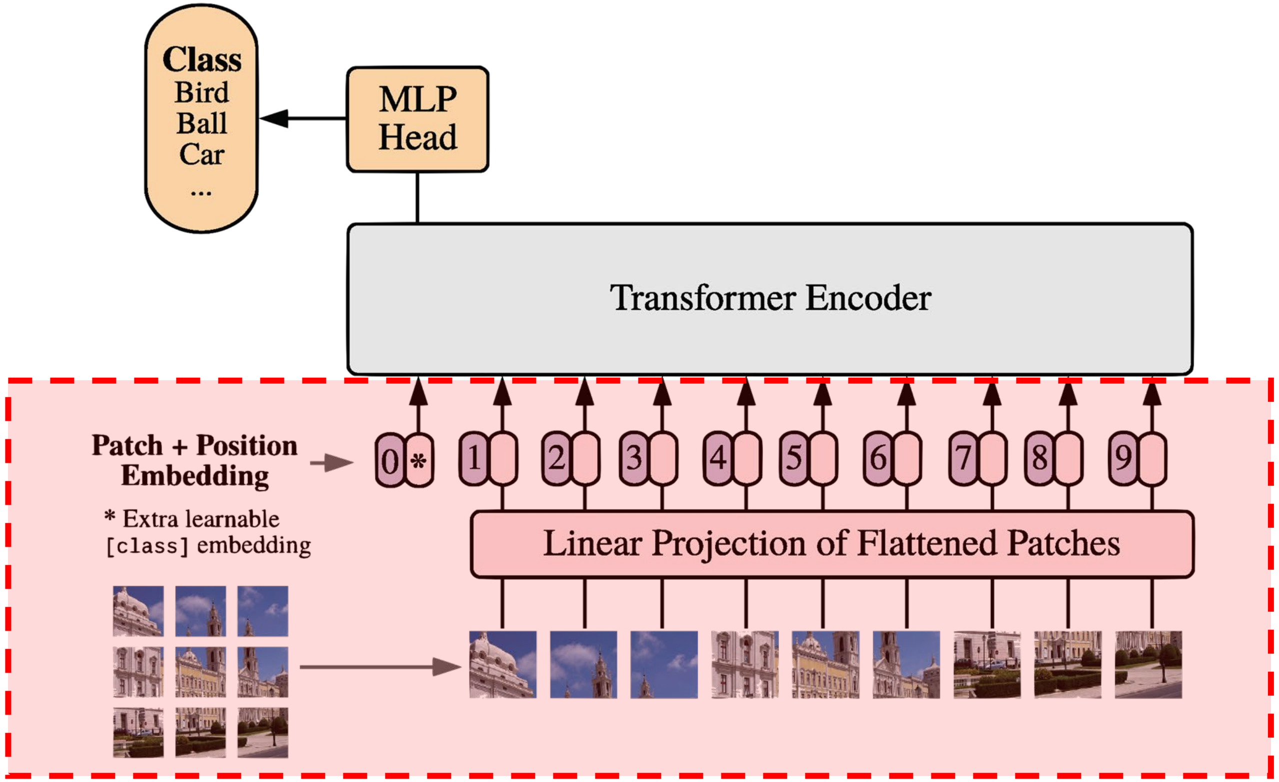
Superpixel Representation

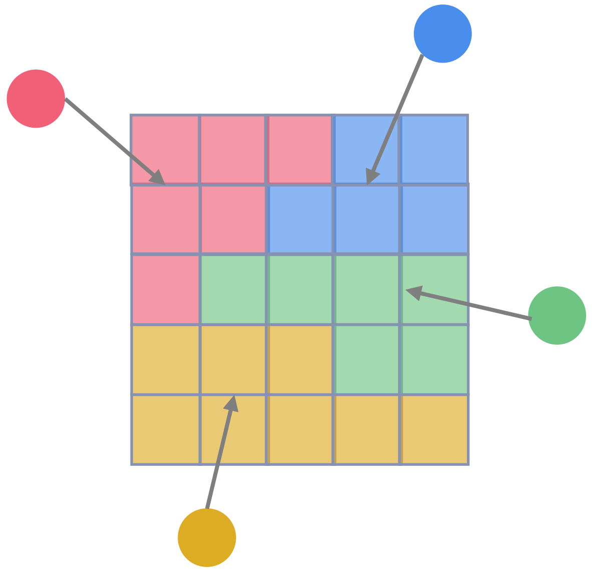
- **Efficient:** lower resolution than pixel/patch
- **Explainable**: formed by grouping pixel features with similar semantics
- **Robust**: rotation & occlusion, driven by Explainability

Patch → Superpixel

It's straight forward to directly **replace the patches by the superpixels**

- However, there are additional challenges





Challenges

Traditional superpixel methods only uses **low level** features (RGB + position)

- Sensitive to low level data augmentation
- Not aware of semantic information

Challenges

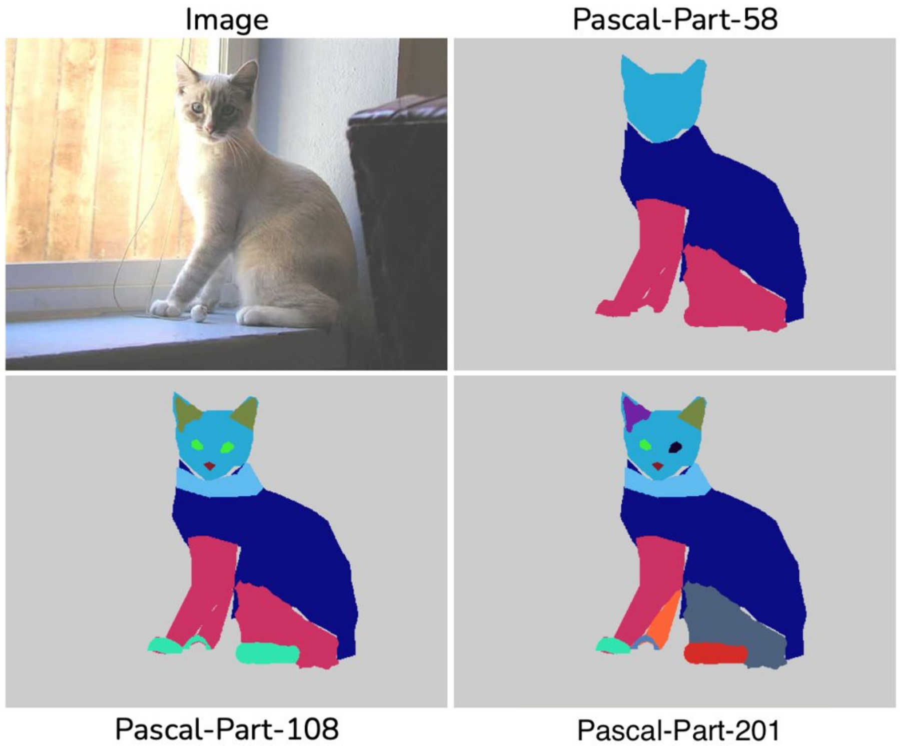
The problem is amplified as

- **Not recoverable** from superpixel errors
- **Not differentiable** due to the hard assignment
    - **Each pixel is assigned to only one superpixel**

Challenges

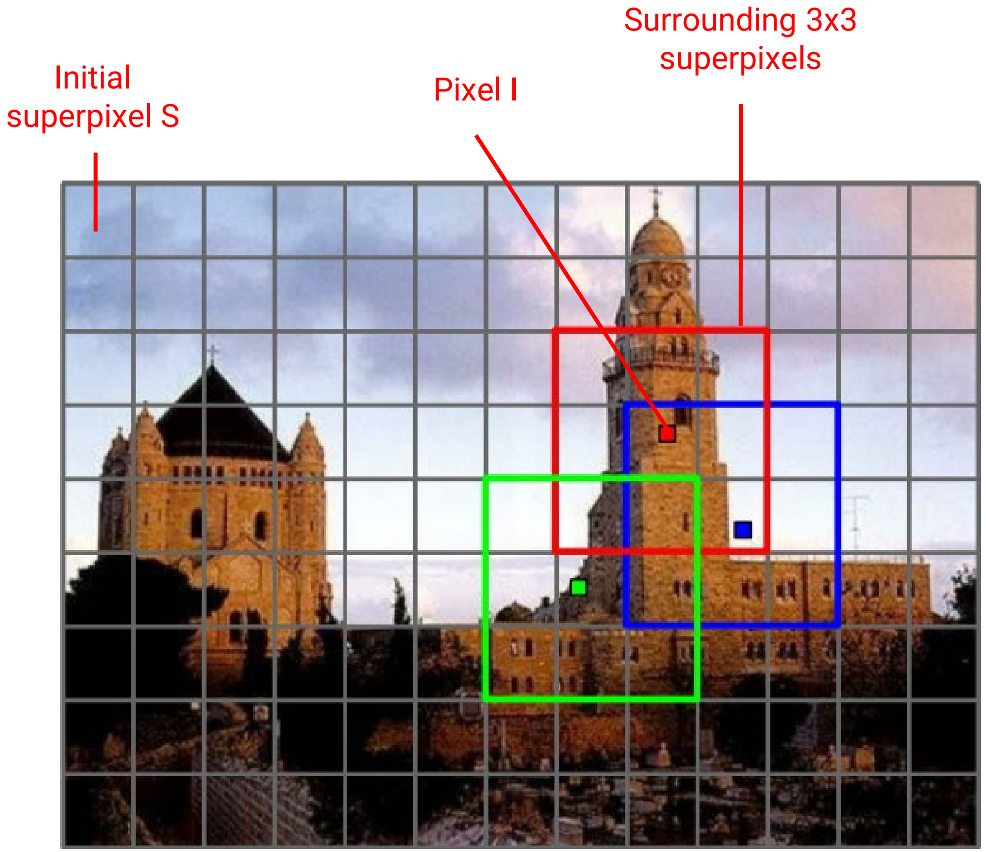
Traditional superpixel method compute a **unique over-segmentation**

- As an over-segmentation method, there is built-in **ambiguity**
- We may require **different granularity** for understanding the image

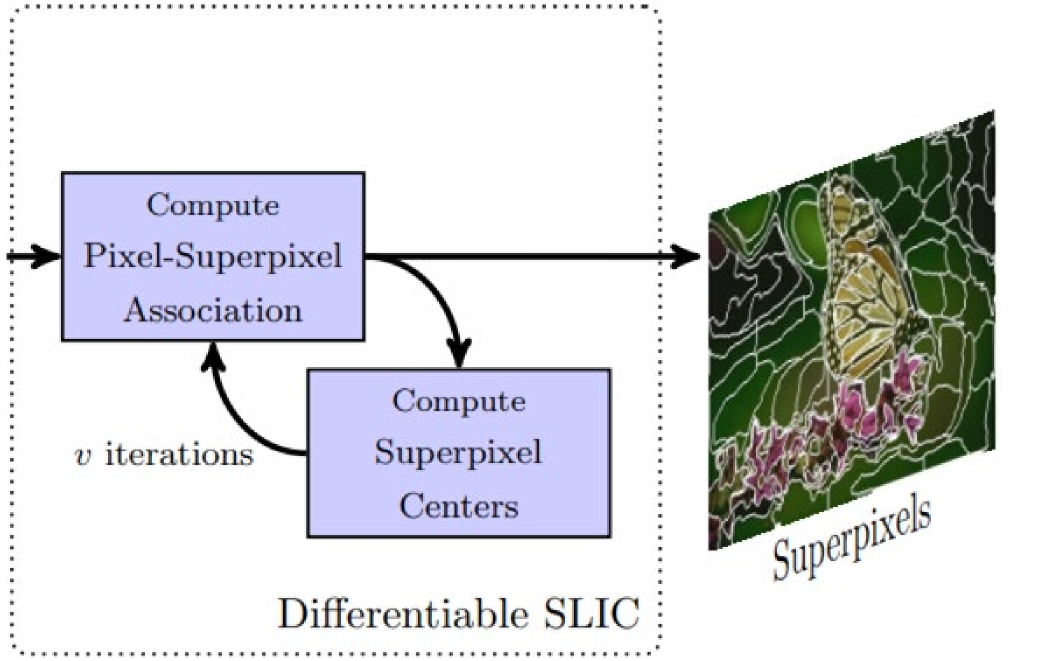


Superpixel as Multi-head Sliding-Window Cross Attention

Preliminary

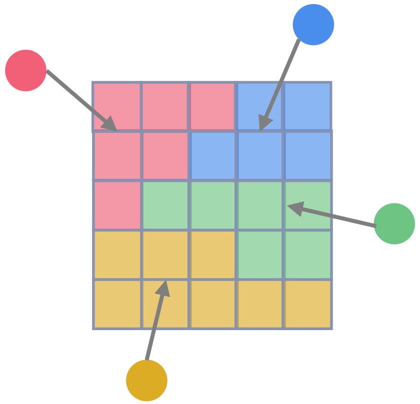


$$Q_{pi}^{t} = e^{-D(I_p, S_i^{t-1})} = e^{-||I_p - S_i^{t-1}||^2}$$



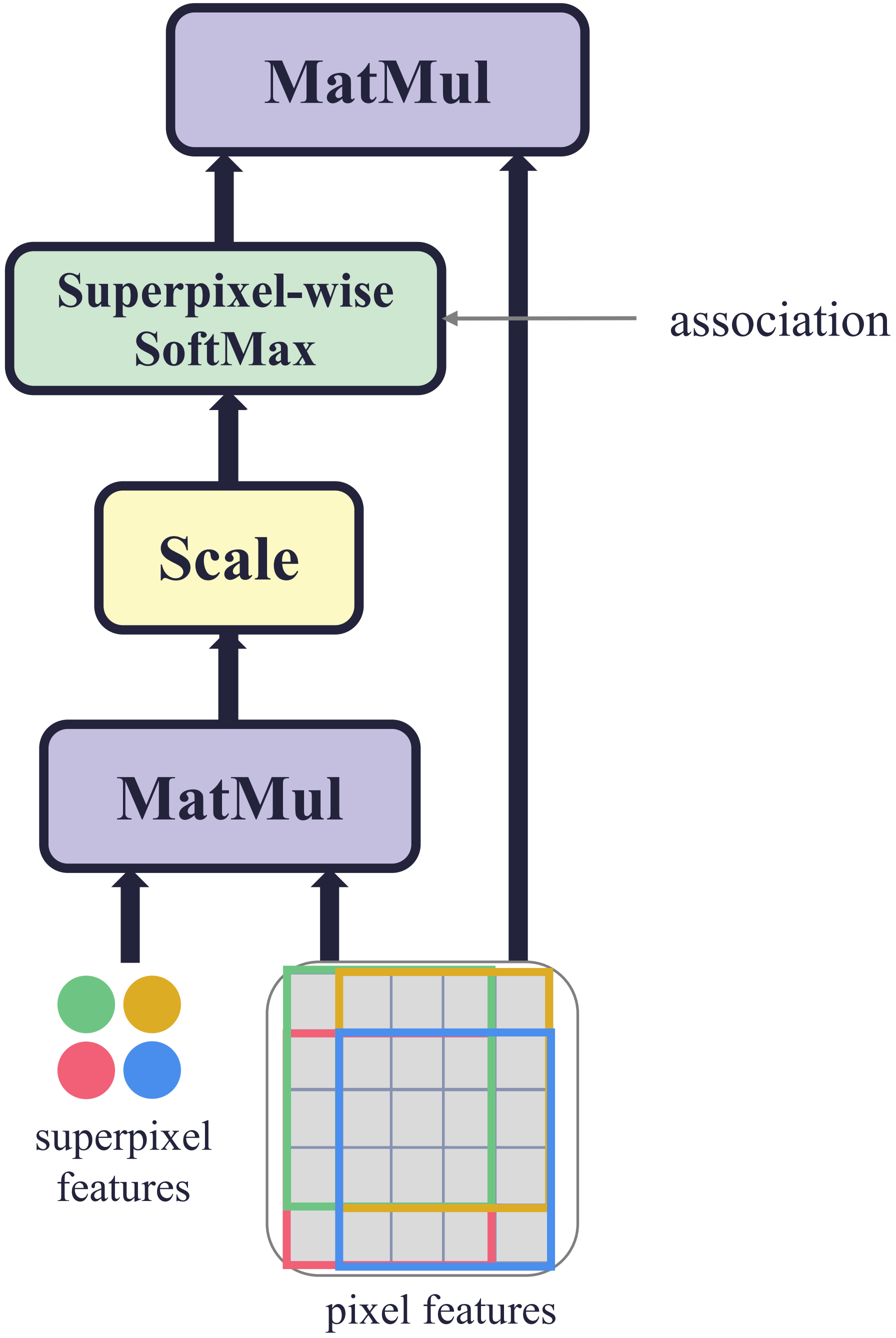
Varun Jampani et al. Superpixel sampling networks. In ECCV, 2018.

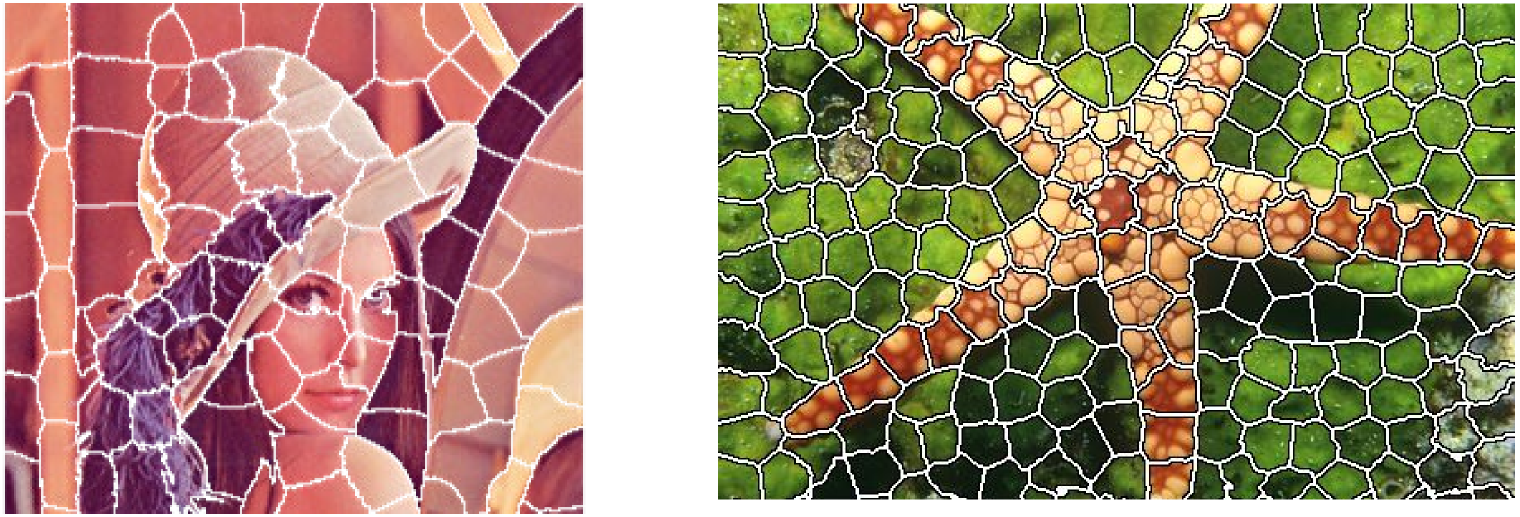
Reformulation



Superpixel Cross Attention

- Multi-head mechanism
  - Multiple superpixel assignment for Ambiguity and granularity
- Superpixel features are updated in a residual manner
  - Ensures the training stability





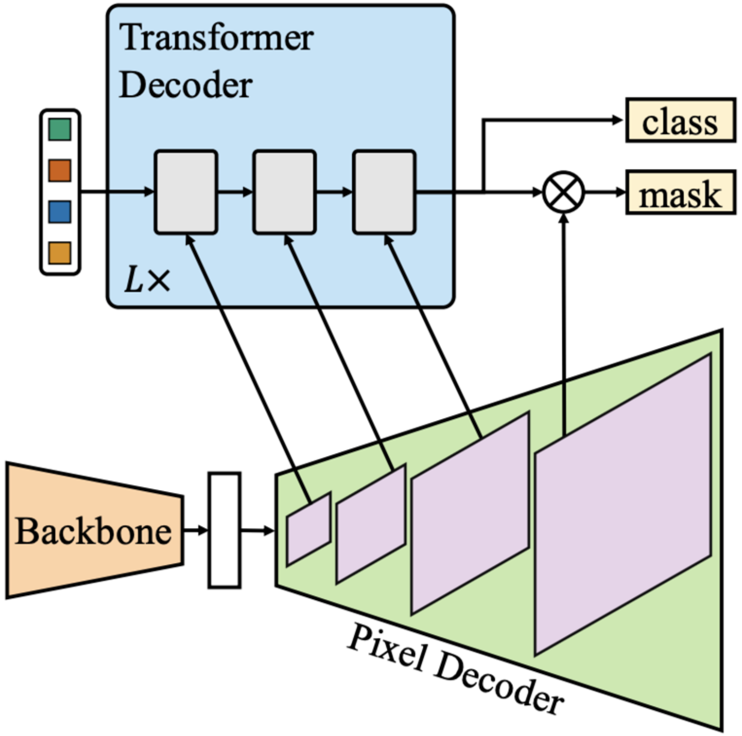
Superpixels will emerge

Superpixel Transformers for Efficient Semantic Segmentation

Motivation

Dense prediction tasks are expensive

- It requires **expensive decoders** which are often stacked upsample-convs.



Bowen Cheng et al. Masked-attention mask transformer for universal image segmentation. In CVPR, 2022.

Motivation

Dense prediction tasks requires **extracting contextual information efficiently**

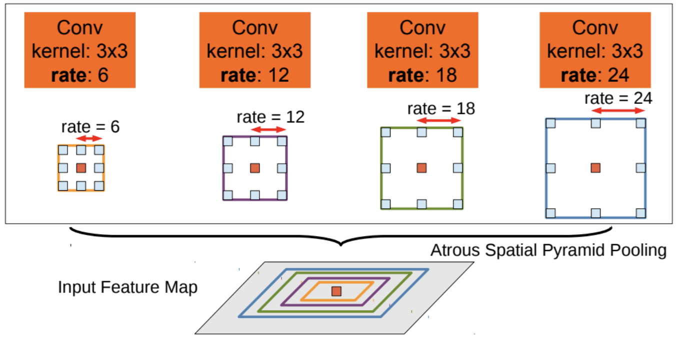
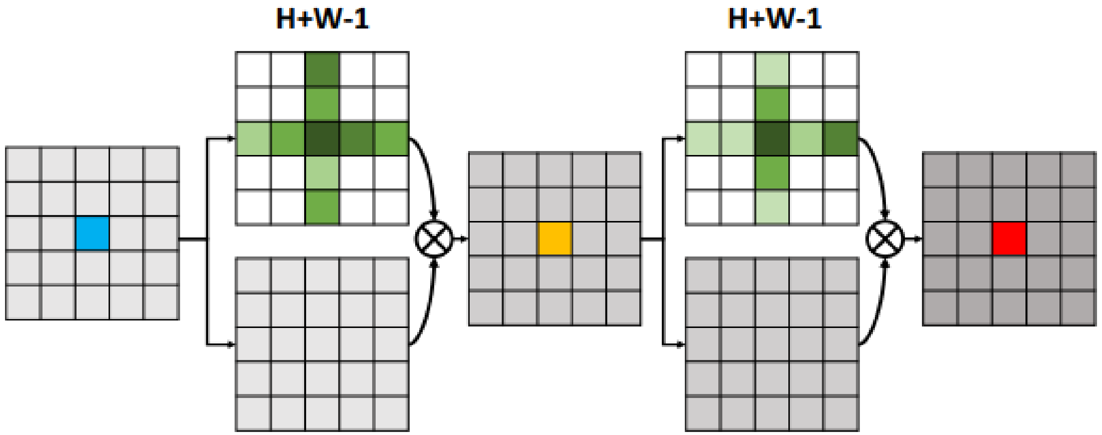


Fig. 4: Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective Field-Of-Views are shown in different colors.



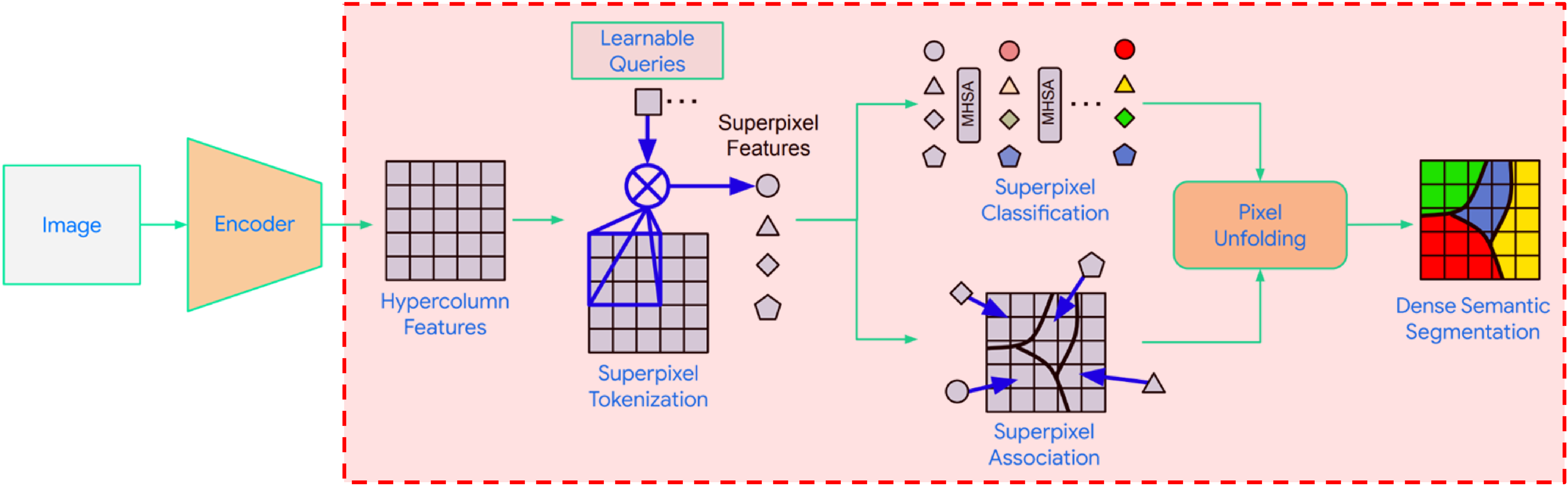
Liang-Chieh Chen et al. in DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. In PAMI, 2018.
Zilong Huang et al. CCNet: Criss-Cross Attention for Semantic Segmentation. In ICCV 2019.

Motivation

- Effectively leverage global context information
- Significantly reduce the computational cost
    - Due to the operation on high resolution pixel features

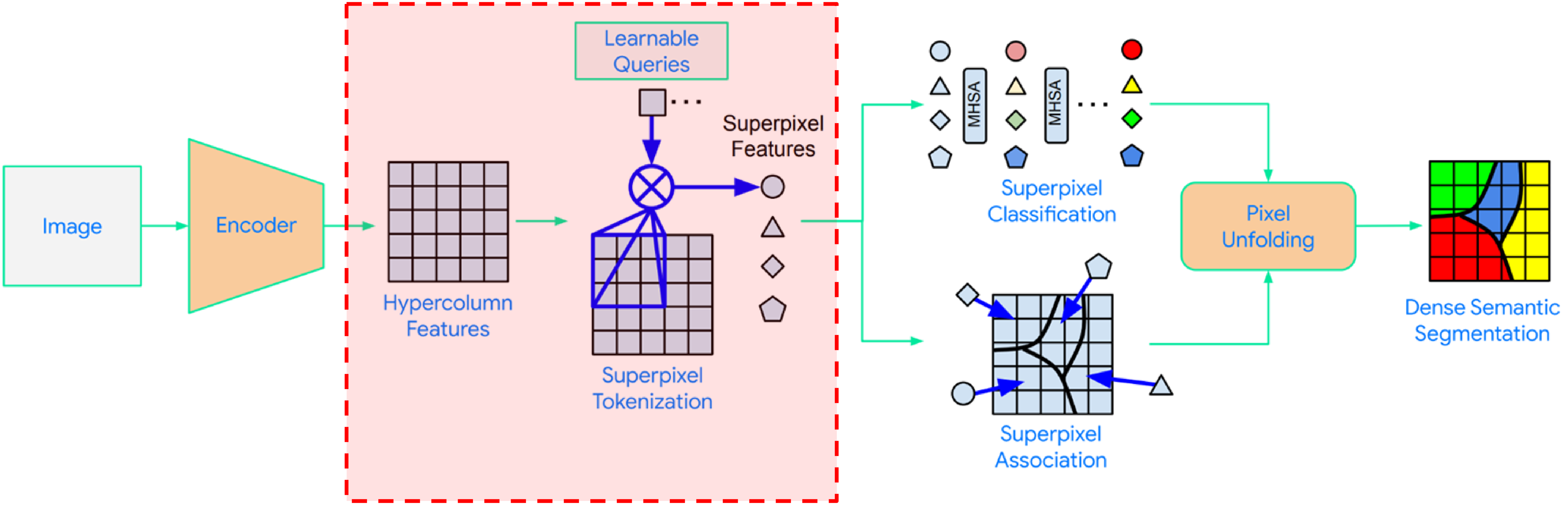
Superpixel Transformer

Replaces the decoder using our superpixel representation



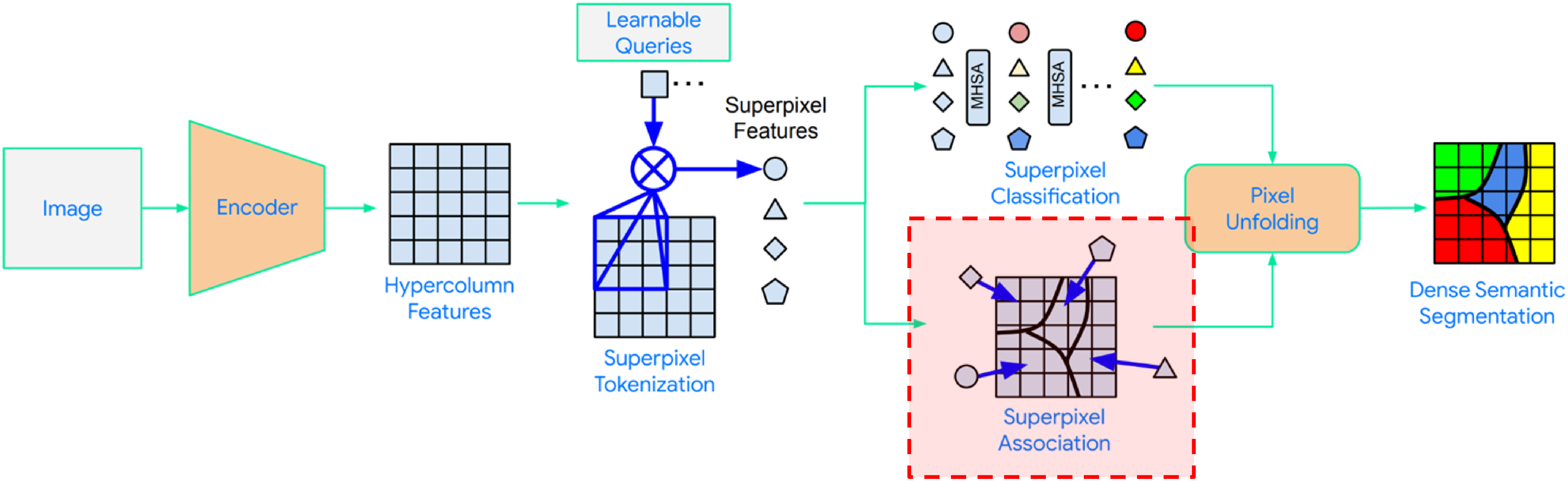
Superpixel Transformer

Extract Superpixel features with reformulation, initialized by learnable queries



Superpixel Transformer

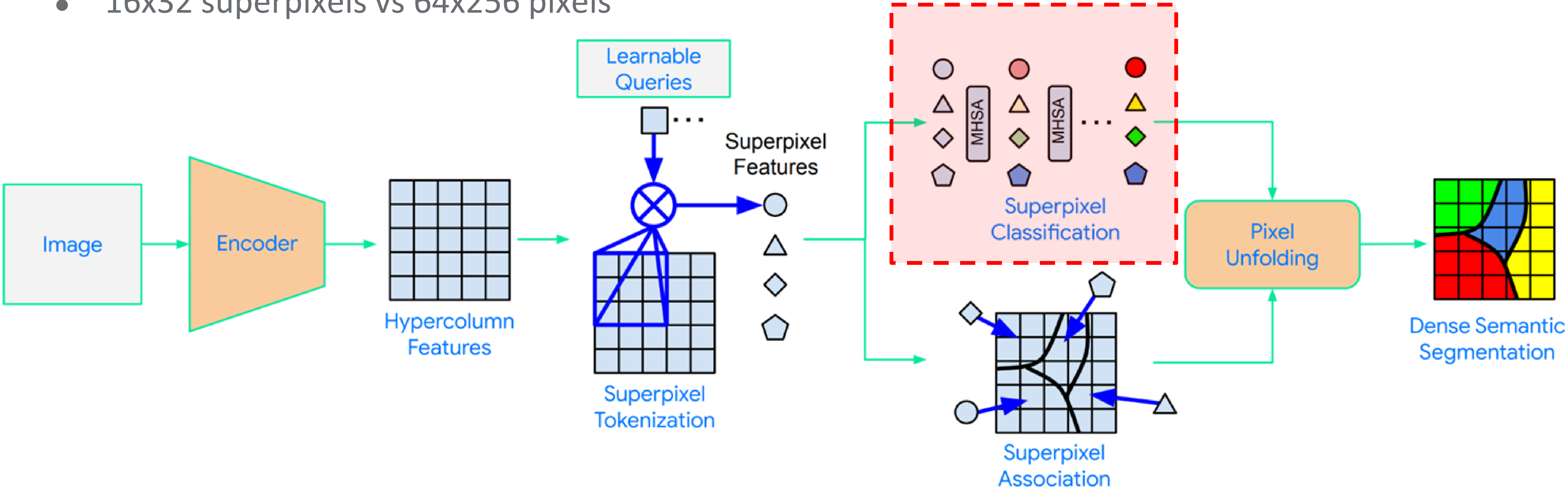
Meanwhile we compute the association between each superpixel and pixel



Superpixel Transformer

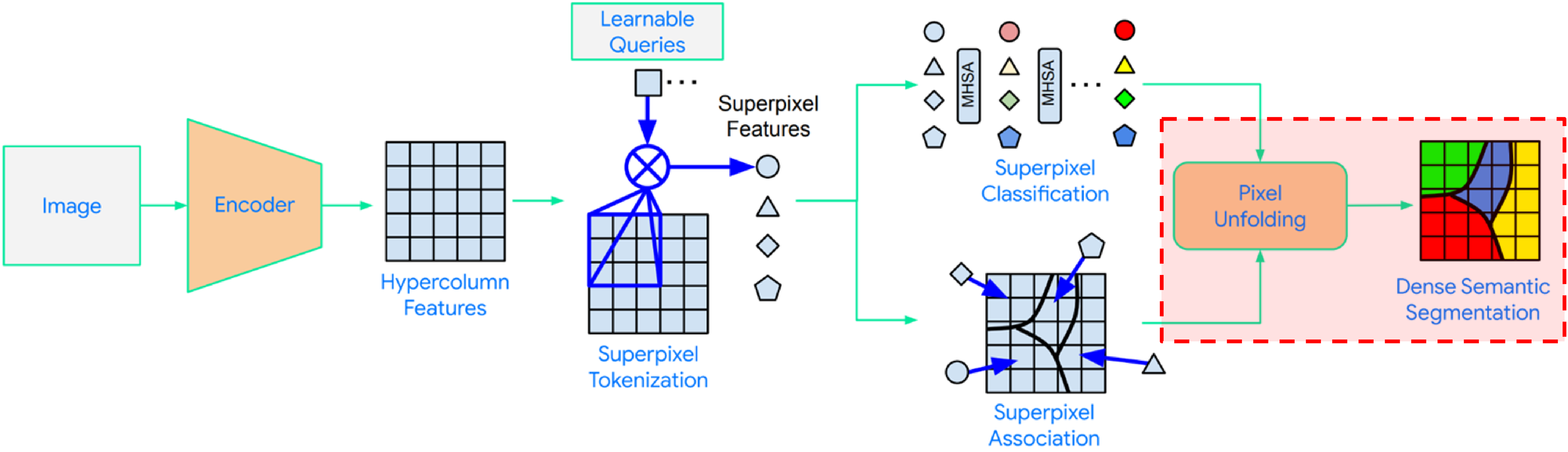
We use global self-attention to enrich the Superpixel features

- 16x32 superpixels vs 64x256 pixels



Superpixel Transformer

Direct **classify each superpixel**
As for each pixel, the final output is a **weighted combination** of the surrounding superpixels' logits, using the **association** instead of bilinear upsampling

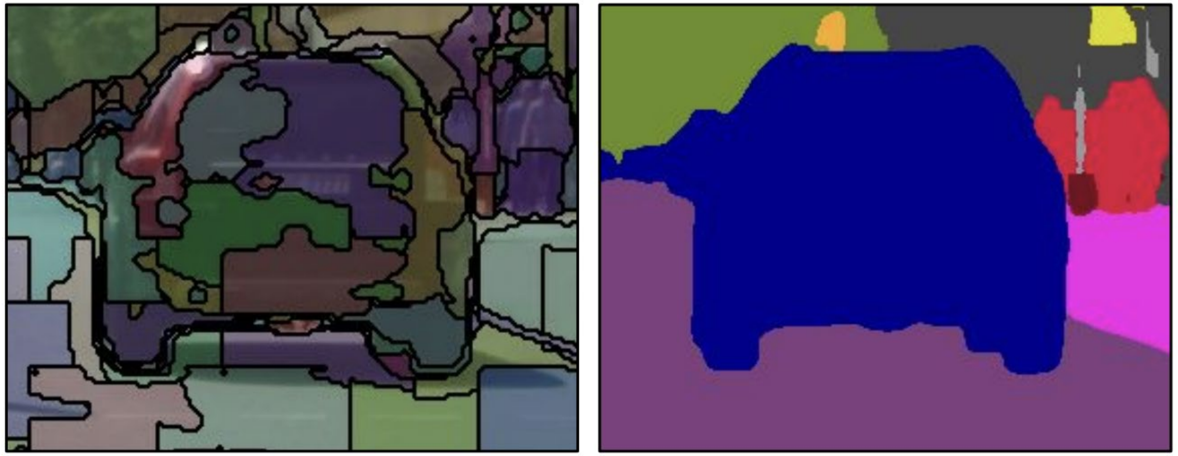


Superpixel Transformer

We decompose the pixel features into a **low dimensional** superpixel representation

By reducing the number of the latent features, we are able to perform **efficient global self-attention** between the superpixel features

Generating the final semantic segmentation predictions is done **entirely by projecting the superpixels** back into the image



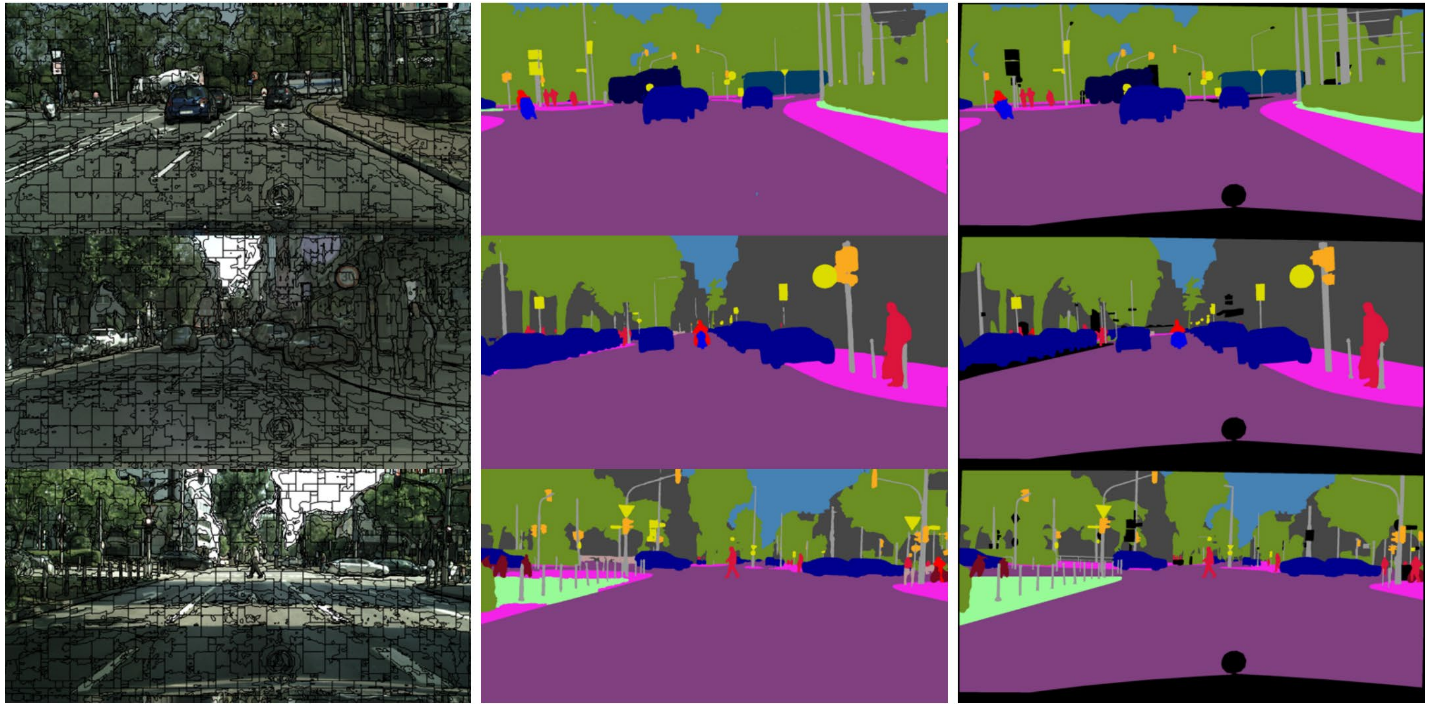
Results on Cityscapes

Method	Backbone	Params ↓	FLOPs ↓	FPS↑	mIoU ↑
MaskFormer [12]	ResNet-50 [21]	-	-	-	78.5
Mask2Former[11]	ResNet-50 [21]	-	-	-	79.4
Panoptic-DeepLab [10]	ResNet-50 [21]	43M	517G	-	78.7
RegProxy* [53]	ViT-S [17]	**23M**	270G	-	79.8
kMaX-DeepLab† [50]	ResNet-50 [21]	56M	434G	9.0	79.7
SP-Transformer	ResNet-50 [21]	29M	**253G**	**15.3**	**80.4**
Mask2Former‡ [11]	Swin-L [31]	-	-	-	83.3
RegProxy* [53]	ViT-L/16 [17]	307M	-	-	81.4
SegFormer [47]	MiT-B5 [47]	**85M**	**1,448G**	2.5	82.4
kMaX-DeepLab† [50]	ConvNeXt-L [32]	232M	1,673G	3.1	**83.5**
SP-Transformer	ConvNeXt-L [32]	202M	1,557G	**3.6**	83.1

Visualization

Argmax of the soft association -> hard assignment

- Can capture thin objects like the poles.
- Shaper edges than the GT
- No direct supervision on the superpixel associations, **implicitly learned** by the network



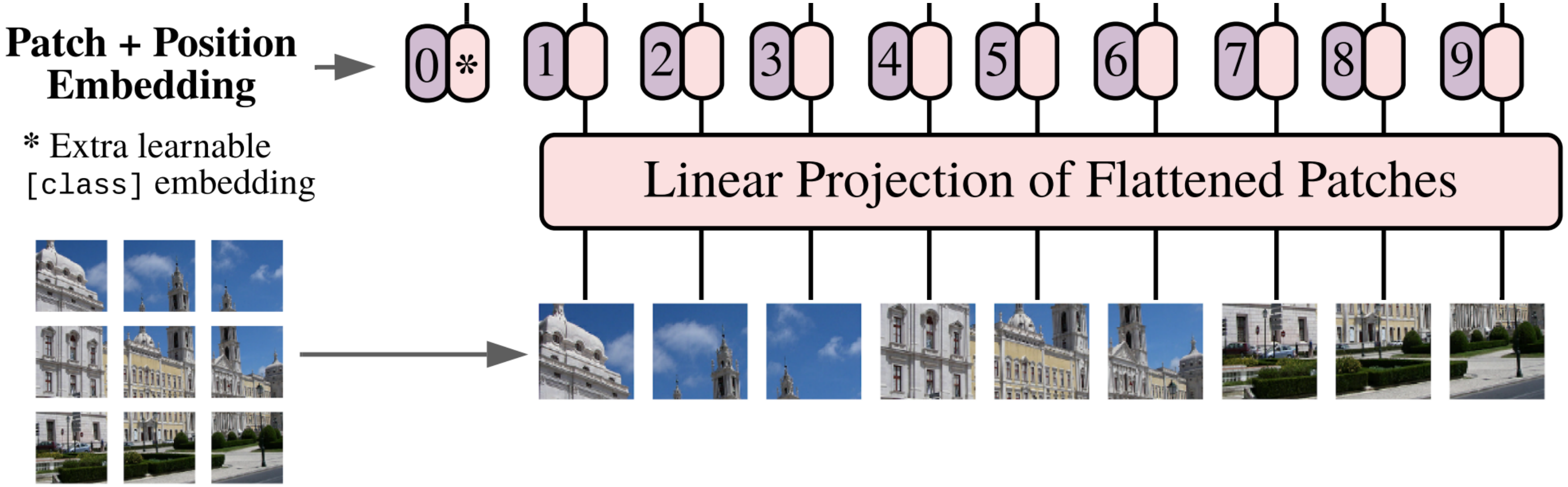
Can we learn superpixels without pixel annotation?

Superpixel Transformer for Classification

Superpixel Embedding

ViT as our baseline

- Pixel features from patch embedding with stride 4
- Superpixel features initialized by 4x4 average pooling

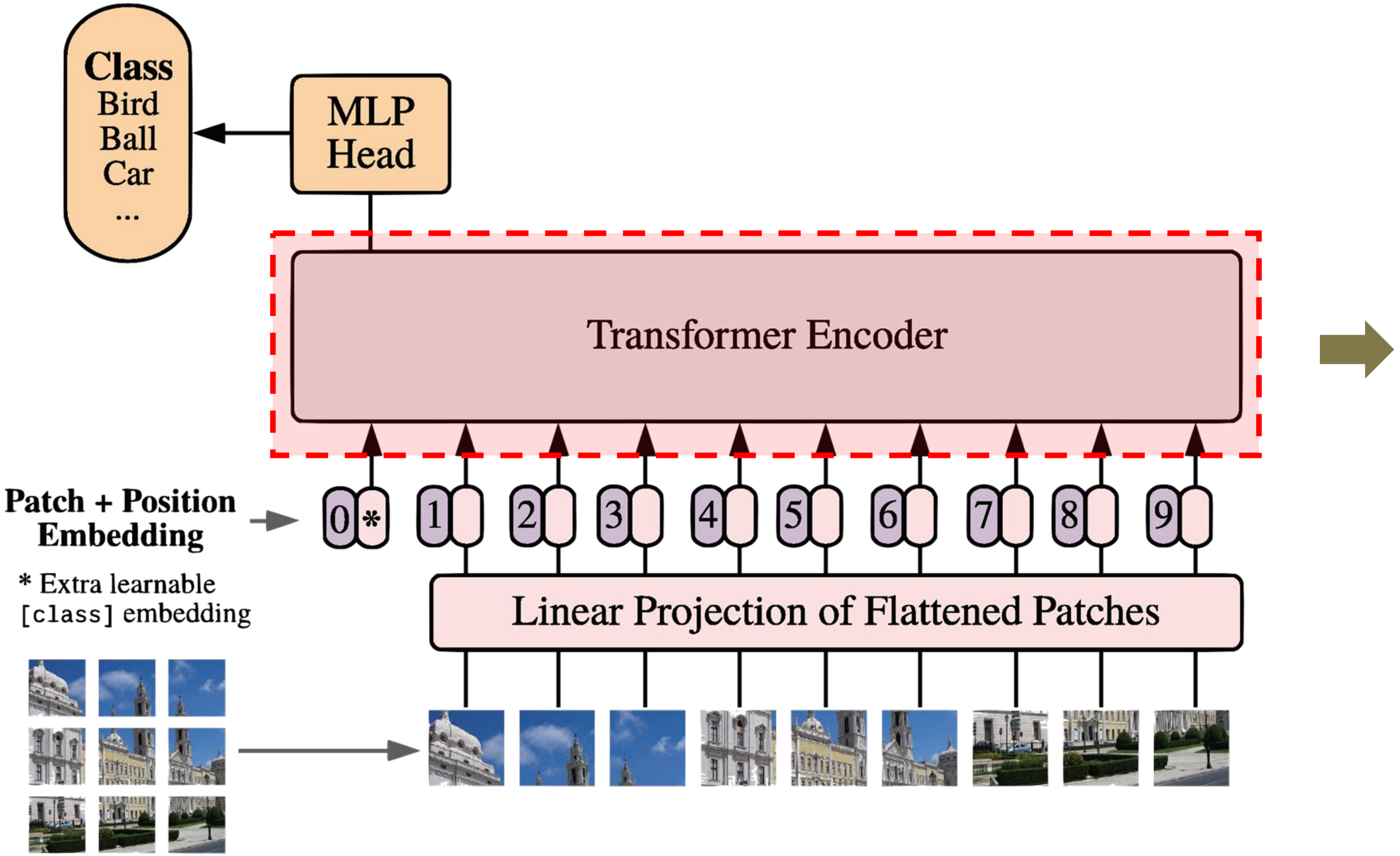


Revisit the Challenges

- Non-differentiability
- Ambiguity & Granularity
- **Superpixels are generated only in low level features**
- **Not recoverable from superpixel errors**

Superpixel Regeneration

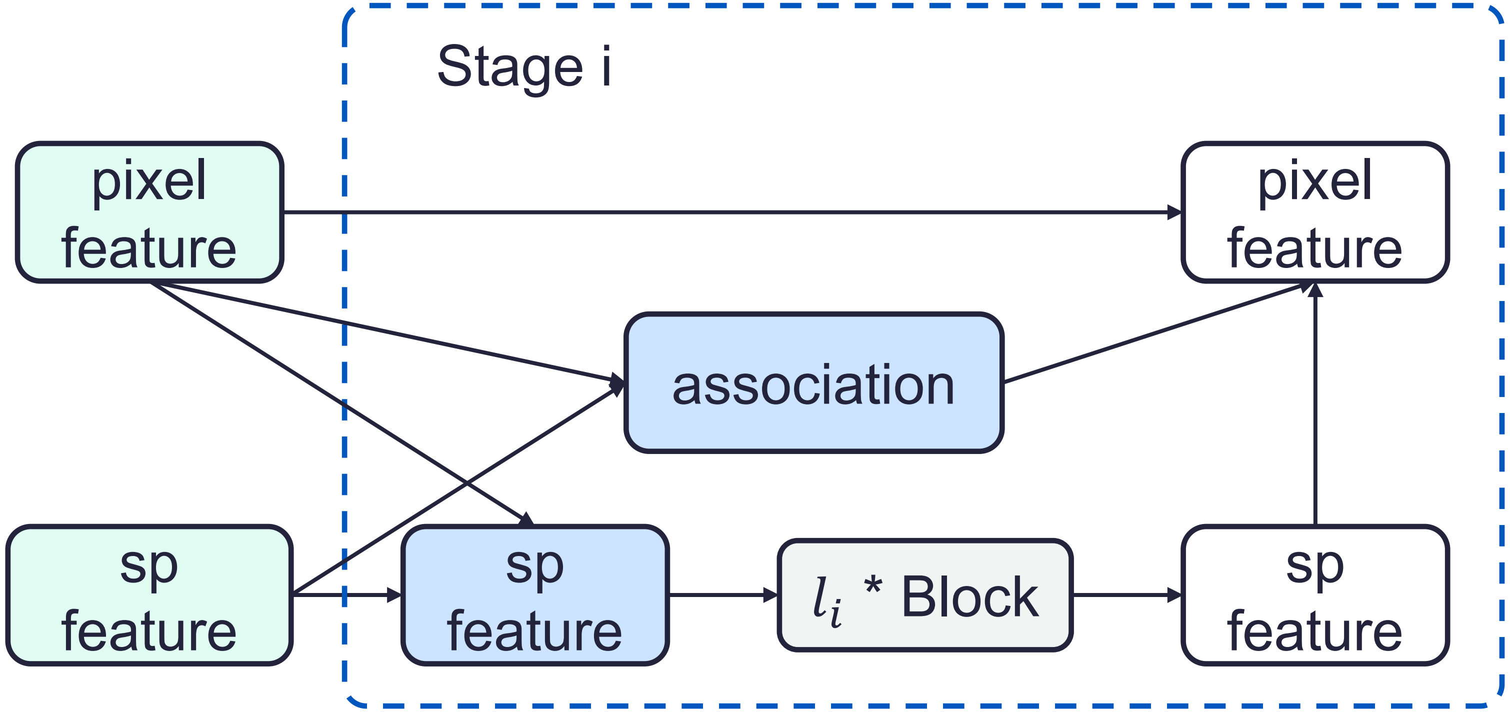
Regenerate the superpixels within the middle of the network using the **enriched features**



Improve the quality of superpixels

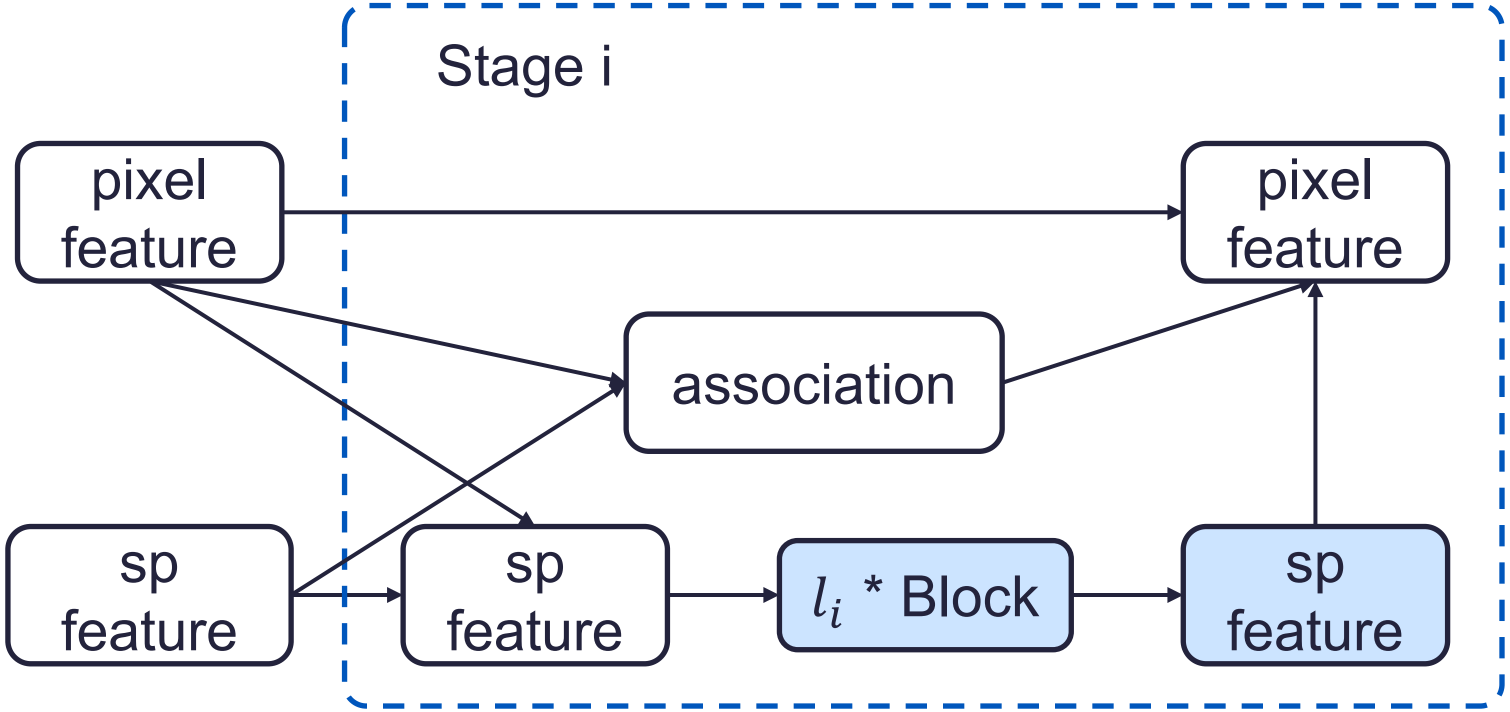
Superpixel Regeneration

We partition the ViT into **multiple stages**



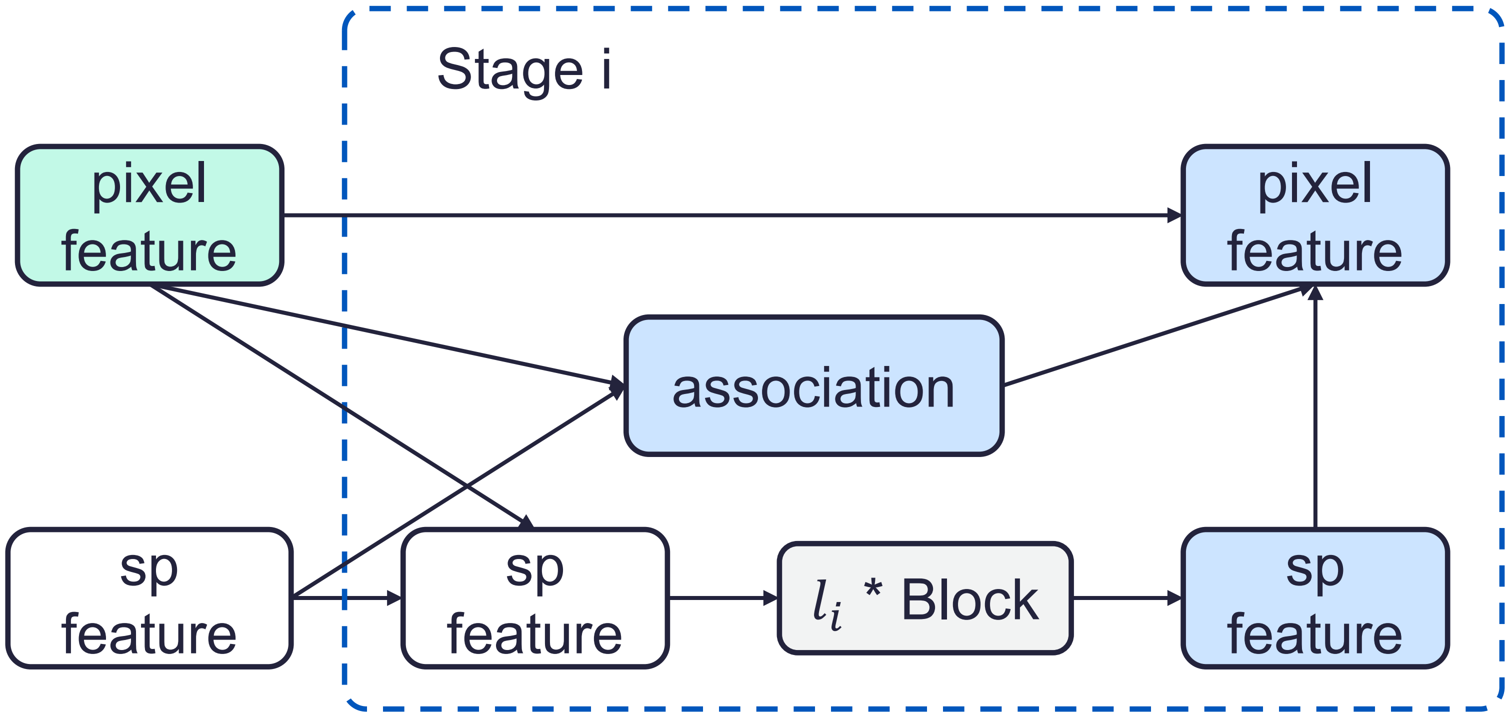
Superpixel Regeneration

Then, we use the **global self-attention** to enrich superpixel features



Superpixel Regeneration

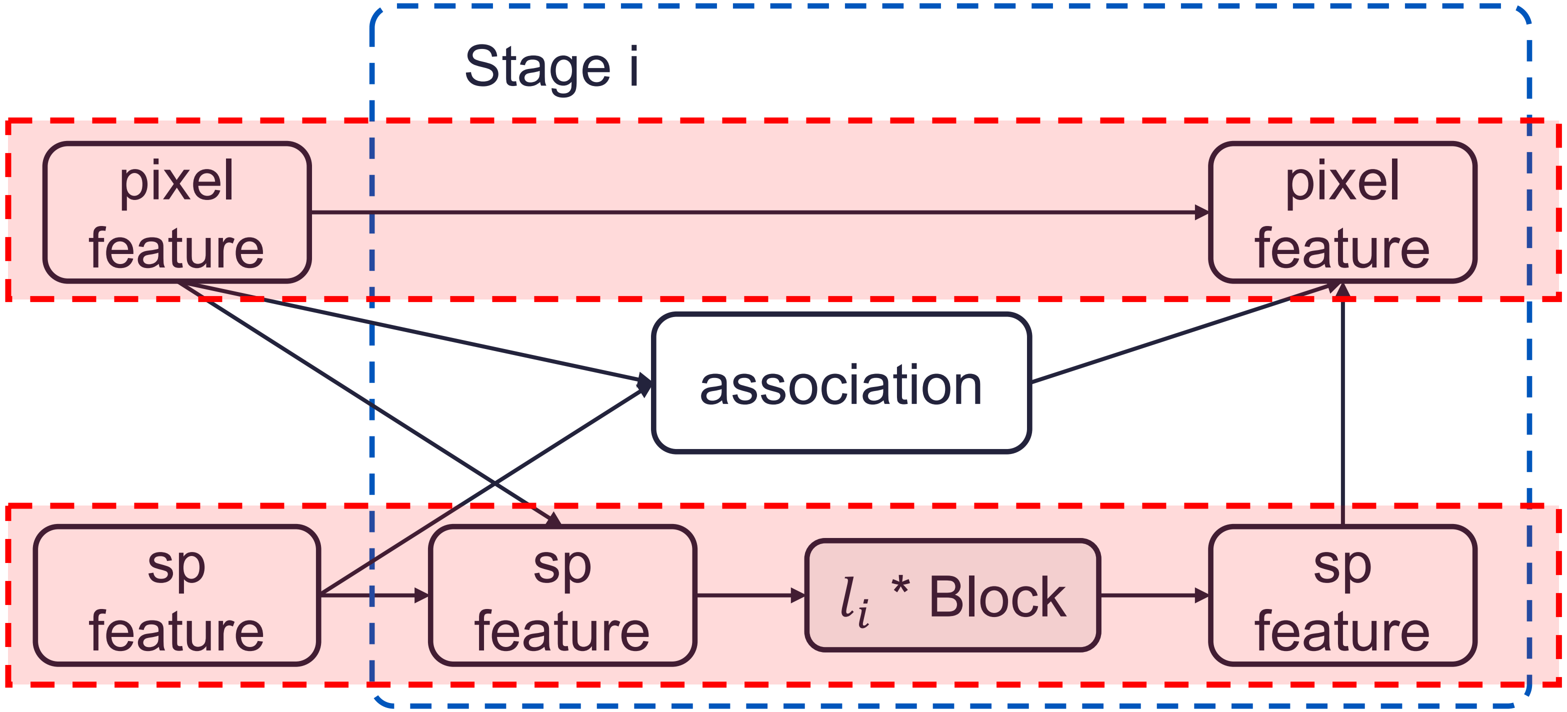
We **enhance the pixel features** through the updated superpixel features, utilize the association for upsampling



Architecture

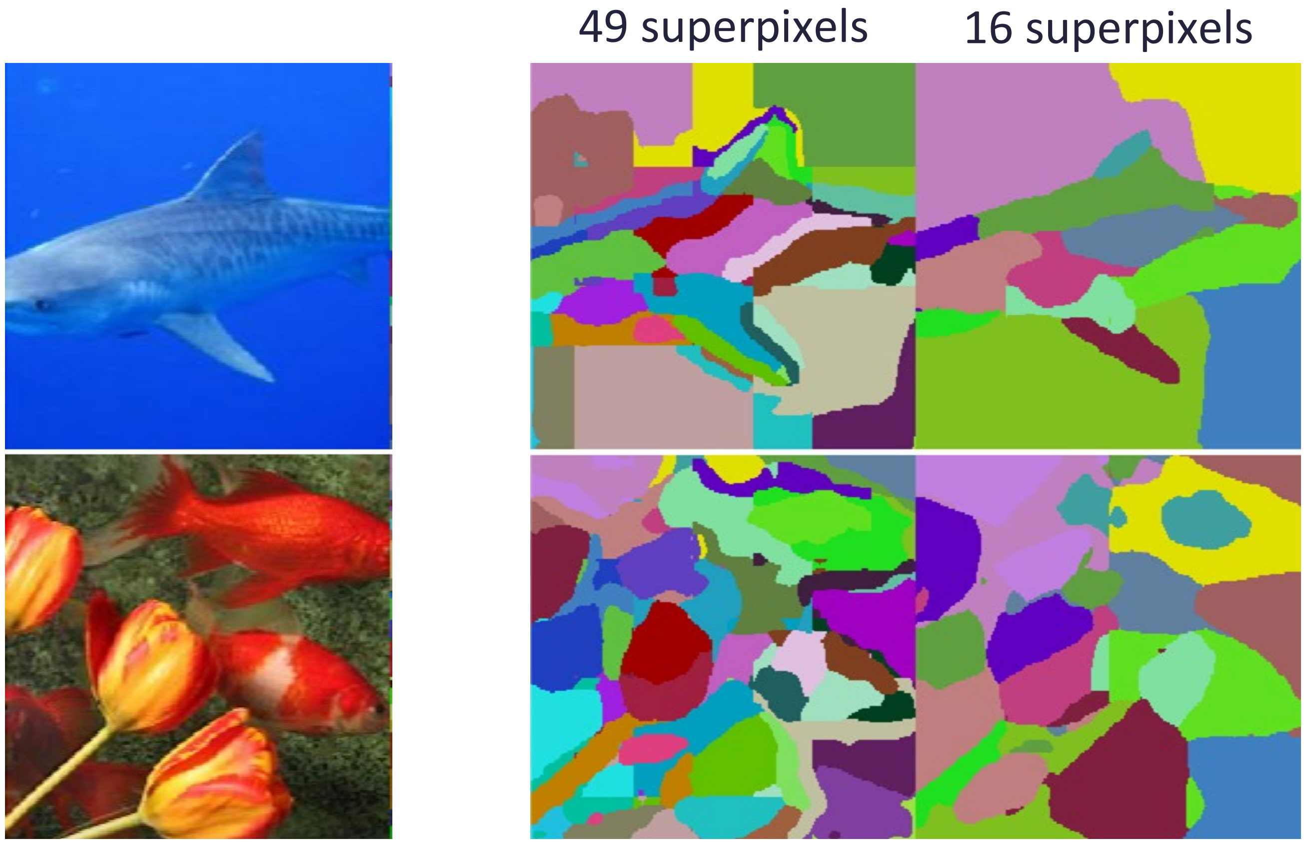
Our network can be viewed as a **dual-branch architecture**.

- Pixel branch: high resolution
- Superpixel branch: low resolution



Visualization

Our method can generate reasonable superpixels even with just **16** tokens



Visualization - Rotation

Our approach produces meaningful superpixels when the image is rotated

Visualization - Occlusion

Our method effectively finds the boundaries and separates them into distinct superpixels

- Patch representation inherently mixes foreground objects with occluders in certain patches

Visualization - Transferability

Our method can transfer to MSCOCO images **without fine-tuning**

COCO

Visualization - Transferability

Our method can produce reasonable results using less superpixels

196 superpixels 49 superpixels 16 superpixels

COCO

Superpixel Quality

Alignment with ground truth boundaries in **Zero-shot** setting

Method	Pascal Voc2012		Pascal-Parts-58	
	mIoU	mAcc	mIoU	mAcc
Patch	87.8	92.8	68.7	78.2
SPFormer-T†	91.5	95.7	71.5	79.9
SPFormer-S†	92.0	96.6	73.3	82.4
SPFormer-B†	91.2	96.3	72.5	81.4
SLIC [1]	92.5	95.4	74.0	81.7

Empirical Results

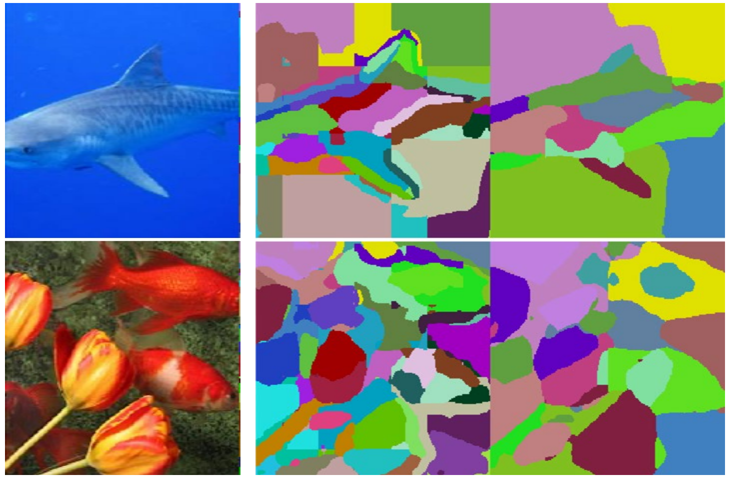
Our models has **much larger capacity**, empowered by the superpixel representation

- Better regularization methods are desired
  - Suitable data augmentation for superpixels?

Method	params	FLOPs	ImageNet Acc.
DeiT-T	5M	1.3G	72.2
SPFormer-T	**5M**	**1.3G**	**73.6 (+1.4)**
DeiT-S	22M	4.6G	79.9
SPFormer-S	**22M**	**4.9G**	**81.0 (+1.1)**
DeiT-Base	87M	17.6G	81.8
SPFormer-B*	**88M**	**18.5G**	**82.4(+0.6)**

*Drop path: 0.1 -> 0.6

Empirical Results



Our method adheres to a **distinct scaling rule** compared to the vanilla ViT

Method	params	FLOPs	ImageNet Acc.
DeiT-S	22M	4.6G	79.9
SPFormer-S	22M	4.9G	81.0 **(+1.1)**
DeiT-S /32	22M	1.1G	73.3
SPFormer-S /32	22M	1.2G	76.1 **(+2.8)**
DeiT-Tiny	5M	**1.3G**	72.2
SPFormer-S /56	22M	**0.5G**	**72.3**

Harness Finer Details

Superpixel representation could benefit from detailed information

Method	params	FLOPs	ImageNet Acc.
DeiT-S	22M	4.6G	79.9
DeiT-S 448	22M	4.6G	80.0 **(+0.1)**
SPFormer-S	22M	4.9G	81.0 **(+1.1)**
SPFormer-S 448	22M	4.9G	81.3 **(+1.4)**

Ablation

Our reformulation mitigates the forementioned challenges

Method	params	FLOPs	ImageNet Acc.
SPFormer-S /32	22M	1.2G	76.1
- multi iterations	22M	1.2G	75.4 **(-0.7)**
- multi stages	22M	1.2G	74.8 **(-1.3)**
- multi head	22M	1.2G	75.6 **(-0.5)**

Superpixel Transformer V2

We successfully learns meaningful superpixels **using only category annotations.**

Leveraging the superpixel representation, our method surpasses the performance of the standard vision transformer, offering **improved efficiency, enhanced explainability, and increased robustness**

From Pixels to Objects: A Hierarchical Approach

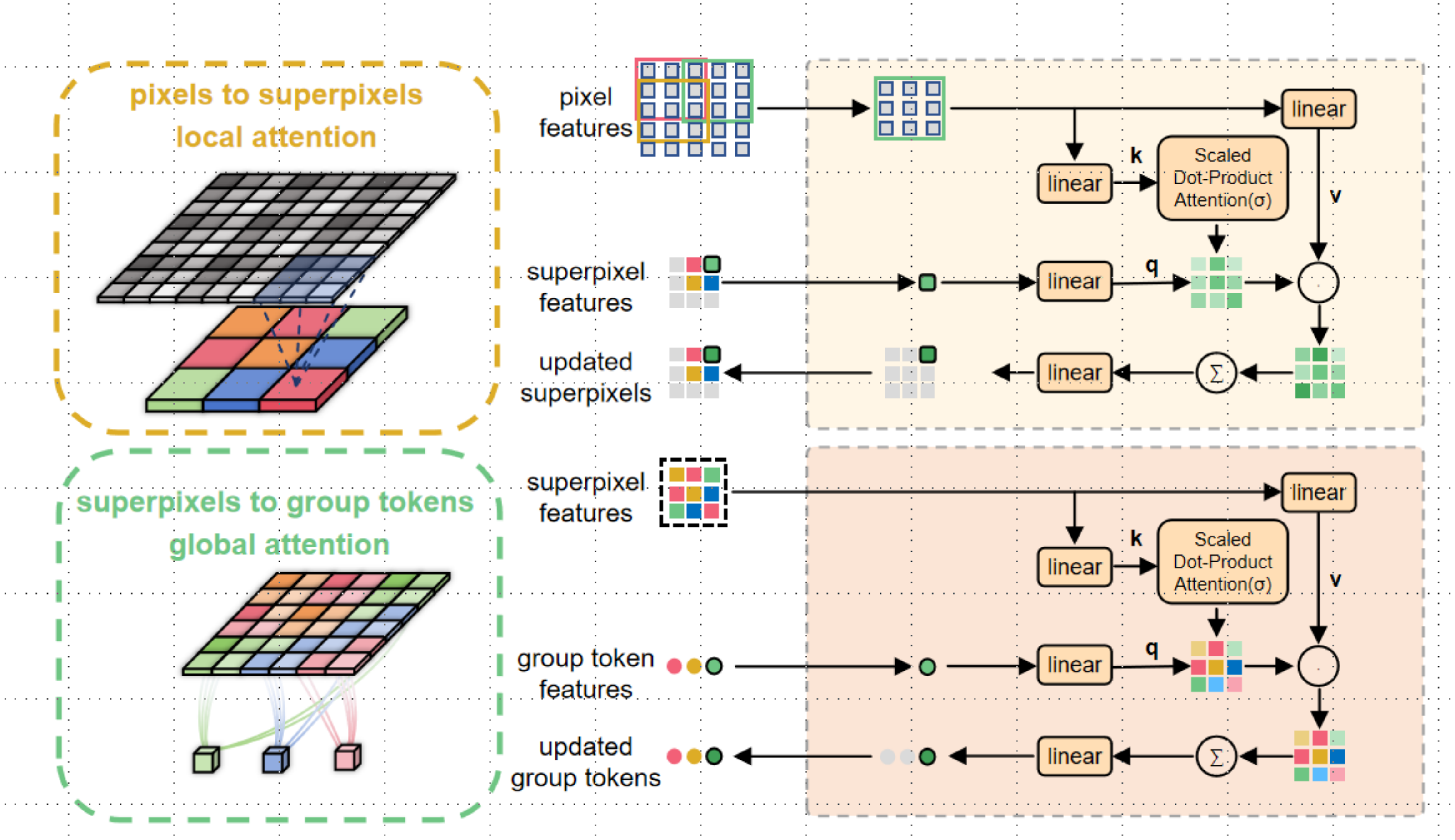
Hierarchy Scene Understanding

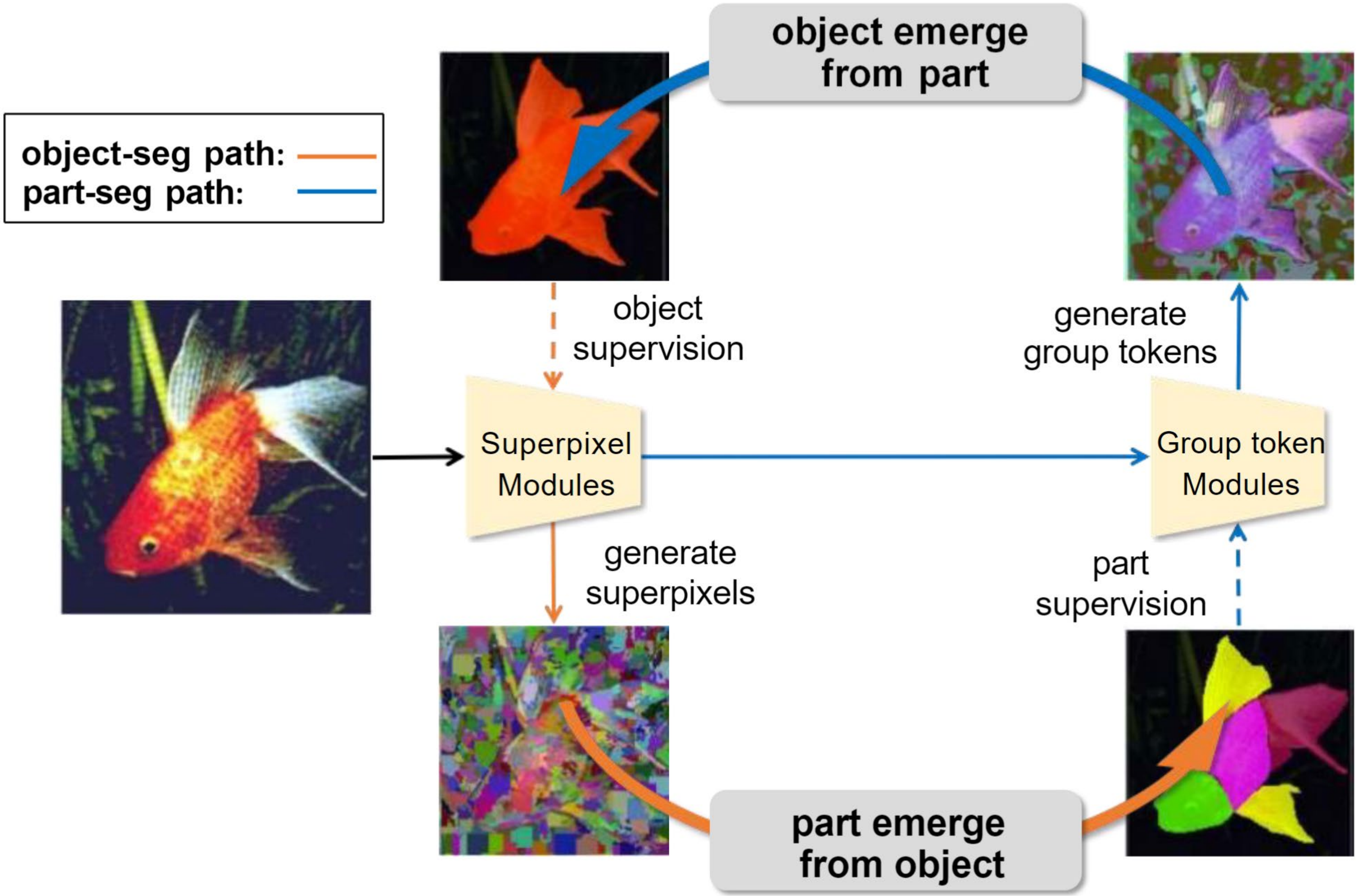
Huge redundancy in areas like the sky

- Arises from the nature of superpixels as an **over-segmentation**
- **Merge similar superpixels** for further increasing the efficiency
    - Through another abstract level: **groups**

Multi-Level Representation

- Pixel -> Superpixel -> Group




object-seg path:
part-seg path:
object emerge from part
object supervision
generate group tokens
Superpixel Modules
Group token Modules
generate superpixels
part supervision
part emerge from object

What’s Next?

Can we get rid of Annotations?

Combine with MAE

MAE couldn't further scale up

- Problem is not hard enough
- One can almost directly copy-paste the pixels to reconstruct the image.

Masking at the superpixel level compels the network **learn to do reasoning**

Supervoxel Transformer

Much more redundancy in 3D

- FLOPs saving: $(\boldsymbol{s}^3)^2$ in 3D
- Combined with the hierarchical scene understanding, we may get **video segmentation & tracking, with or without annotations**.

Chenliang Xu et al. Evaluation of Super-Voxel Methods for Early Video Processing. In CVPR, 2012.