Context Understanding

3 Visual Transformer

Context Understanding
Visual Transformer

P. Ramachandran et al. Stand-alone self-attention in vision models. NeurIPS

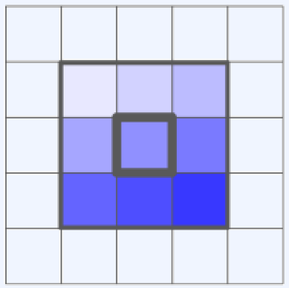


Figure 1: An example of a local window around $i = 3, j = 3$ (one-indexed) with spatial extent $k = 3$.

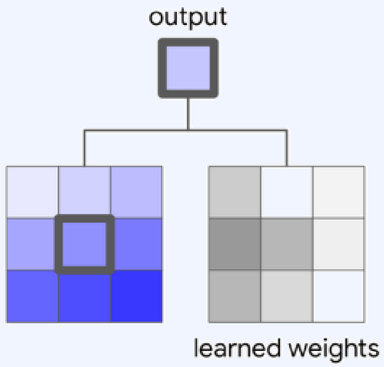
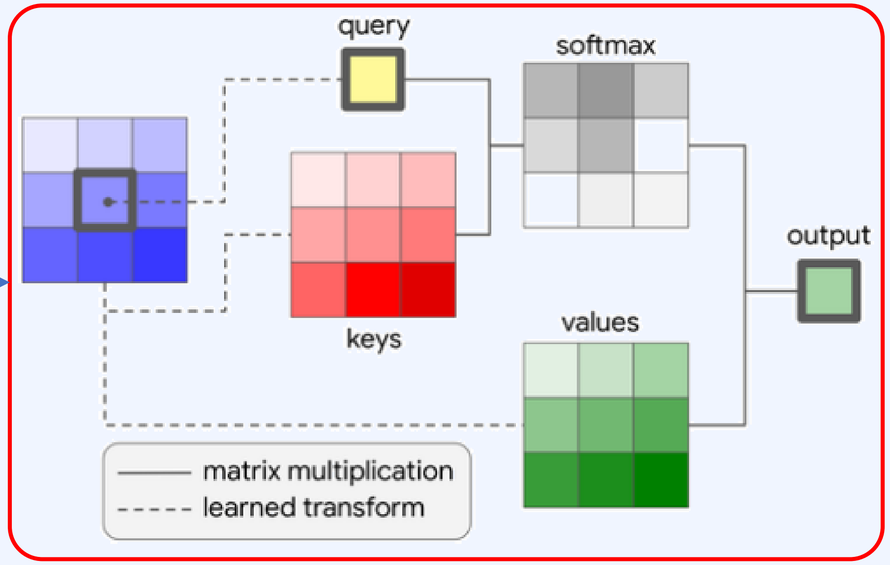


Figure 2: An example of a 3×3 convolution. The output is the inner product between the local window and the learned weights.

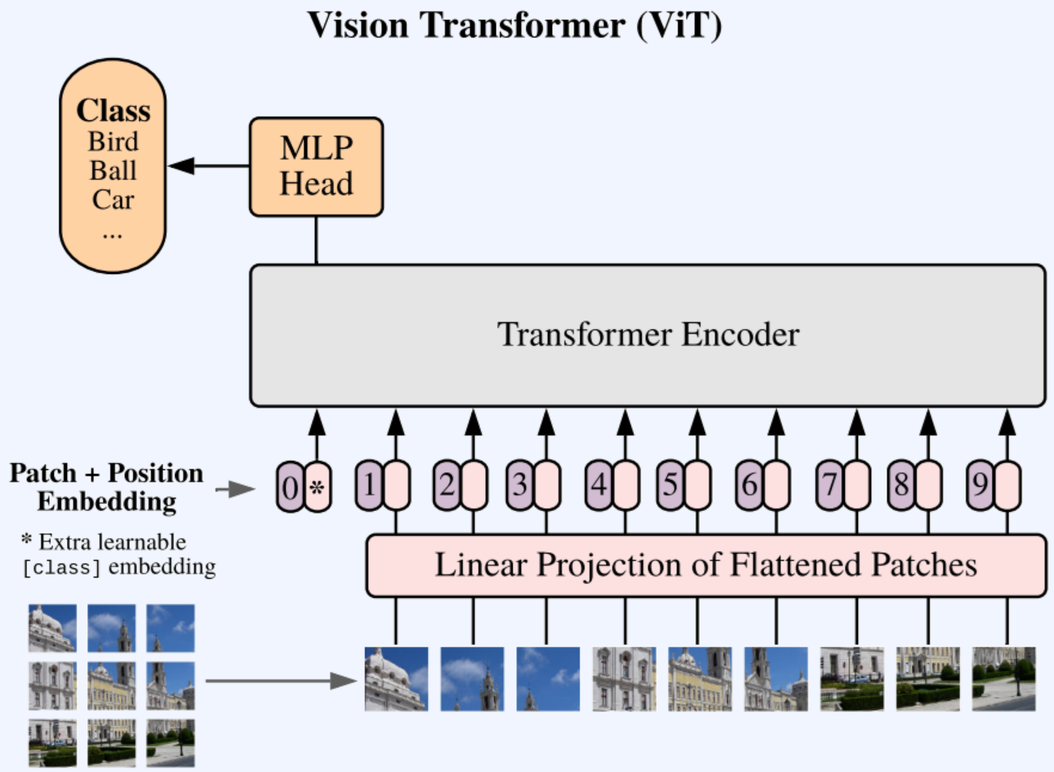


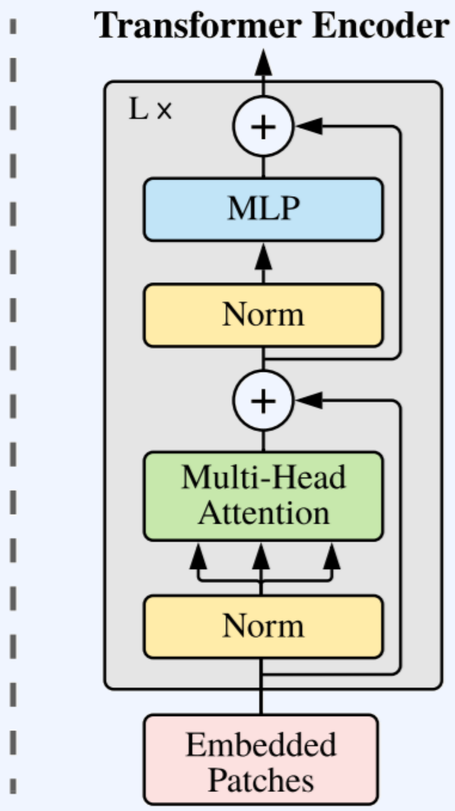
-1, -1	-1, 0	-1, 1	-1, 2
0, -1	0, 0	0, 1	0, 2
1, -1	1, 0	1, 1	1, 2
2, -1	2, 0	2, 1	2, 2

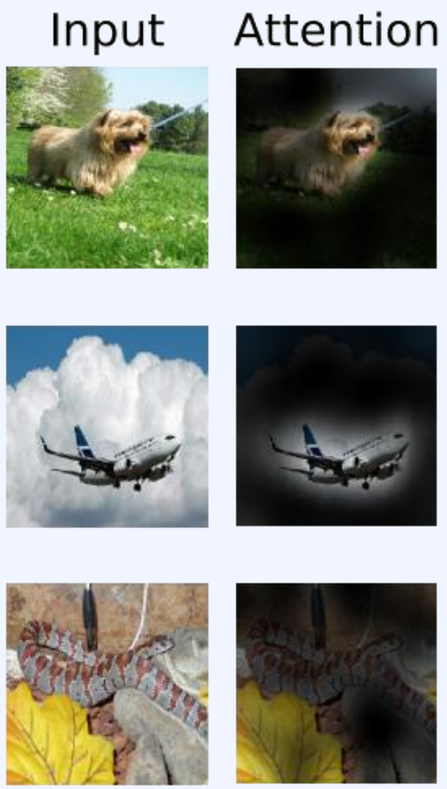
Relative distance computation

Context Understanding
Visual Transformer

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR

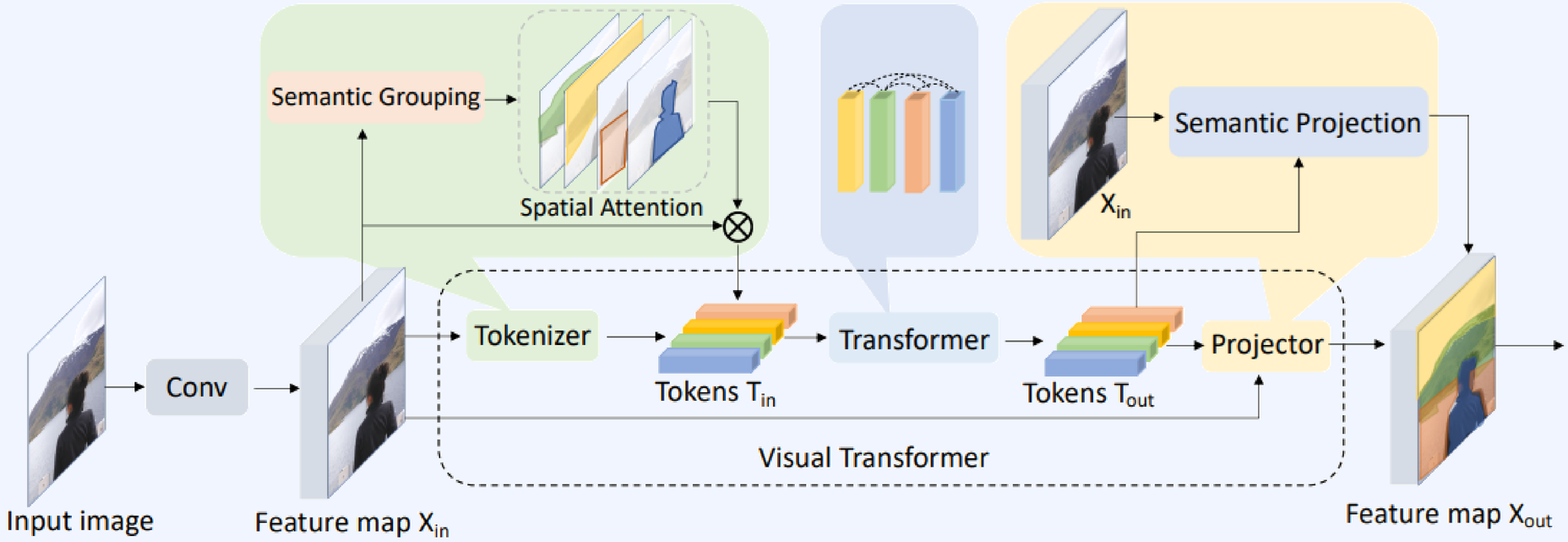


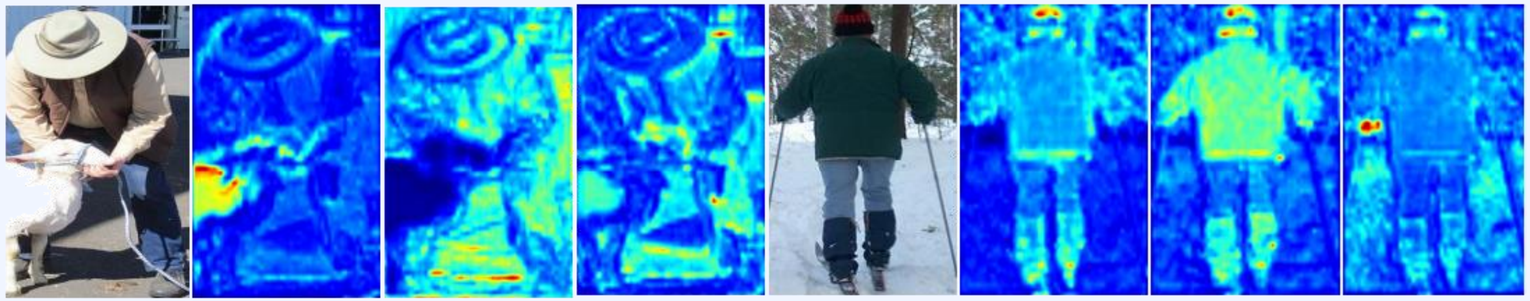




Context Understanding
Visual Transformer

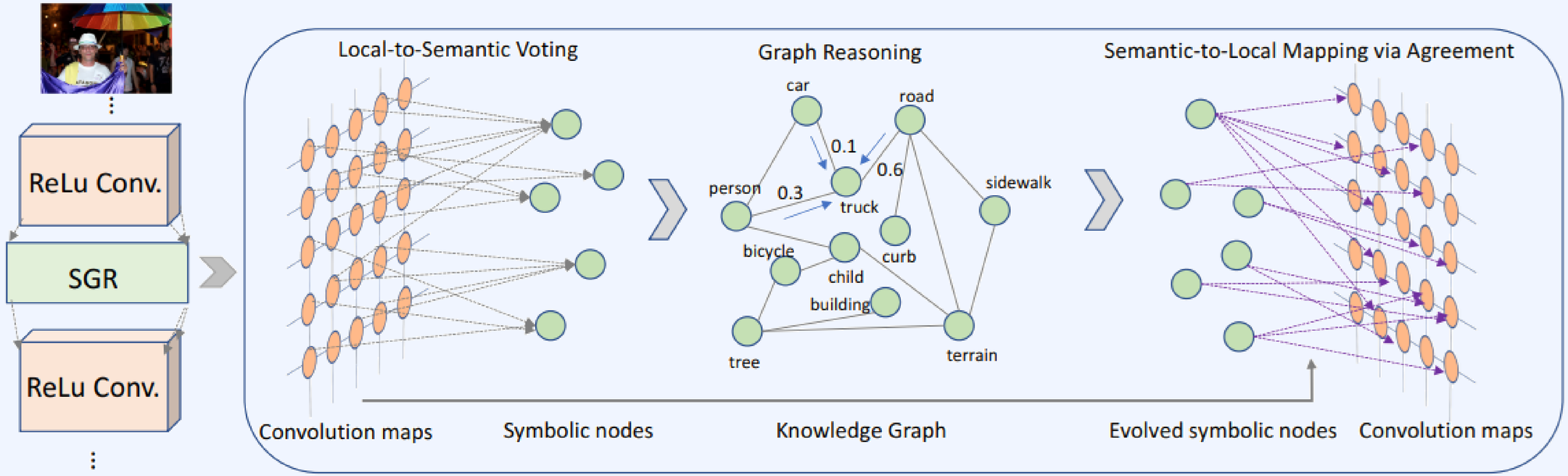
B. Wu et al. Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677





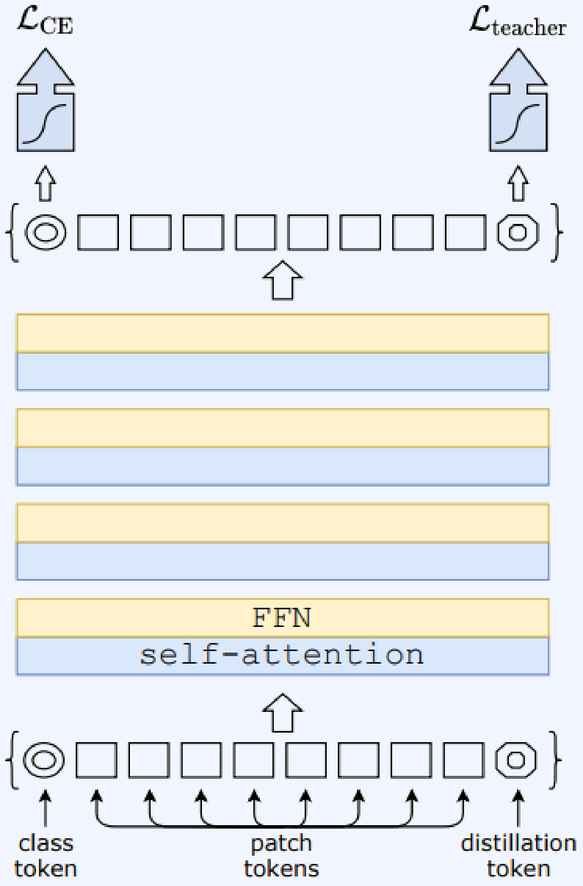
Context Understanding
Visual Transformer

X. Liang et al. Symbolic Graph Reasoning Meets Convolutions. NeurIPS

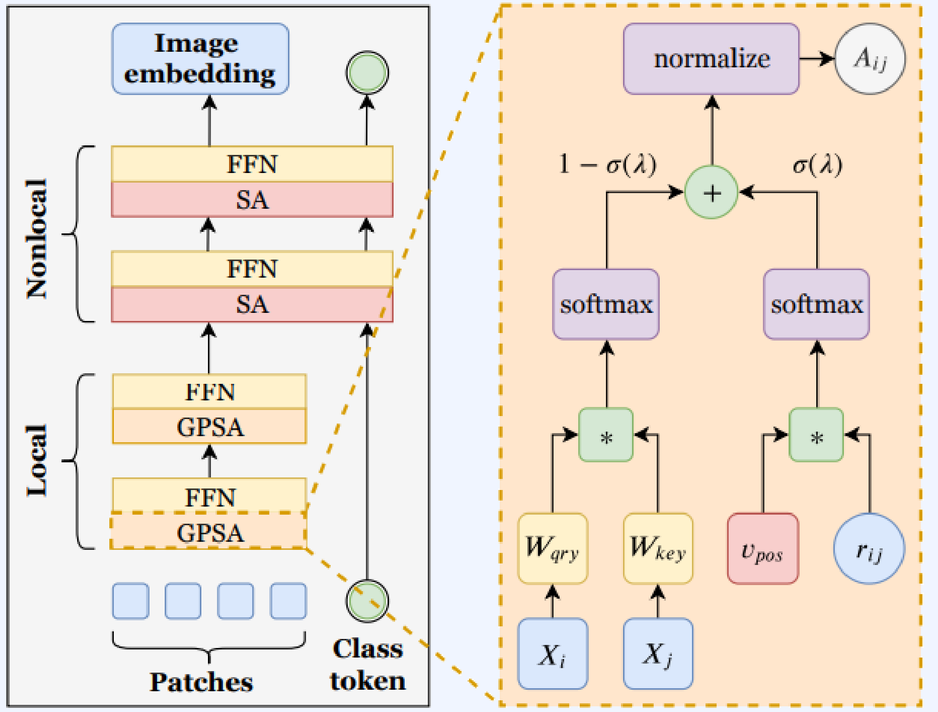


Context Understanding
Visual Transformer

H. Touvron et al. Training data-efficient image transformers & distillation through attention. ICLR



DeiT
Teacher (CNN) – Student (ViT)



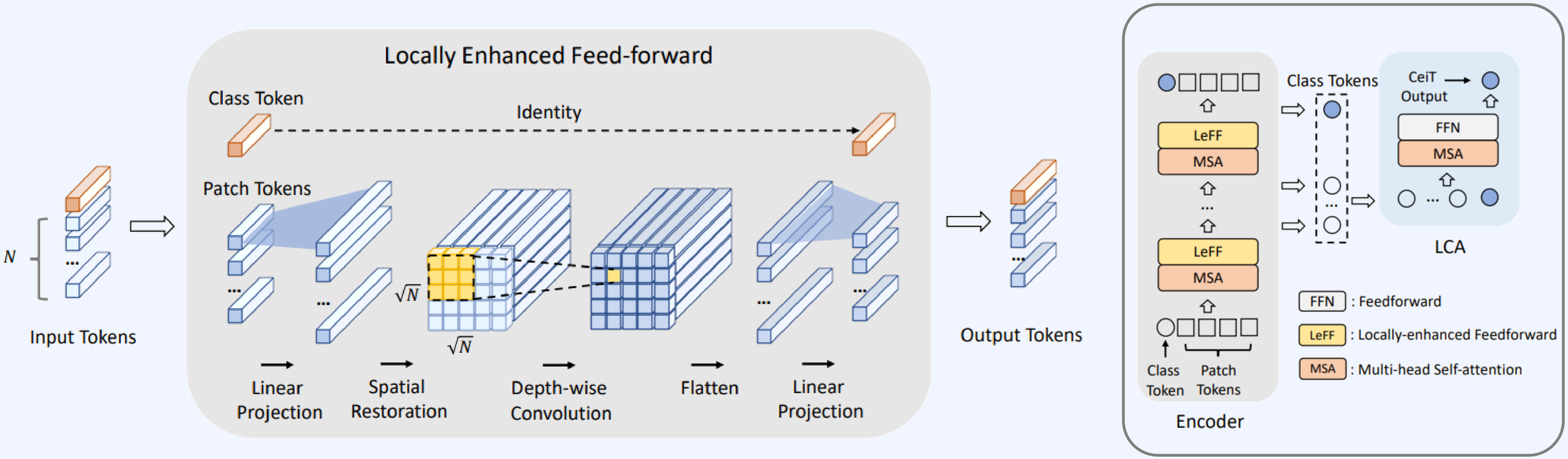
ConViT **Gated Positional Self-Attention**

S. d'Ascoli et al. Convit: Improving vision transformers with soft convolutional inductive biases. ICLR

Context Understanding
Visual Transformer

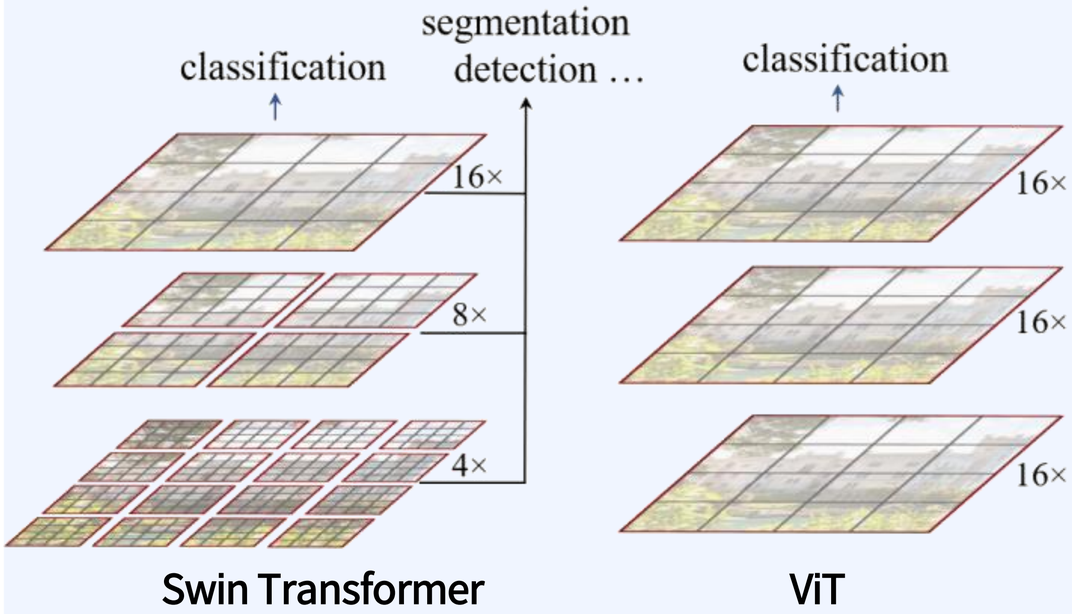
K. Yuan et al. Incorporating convolution designs into visual transformers. In ICCV

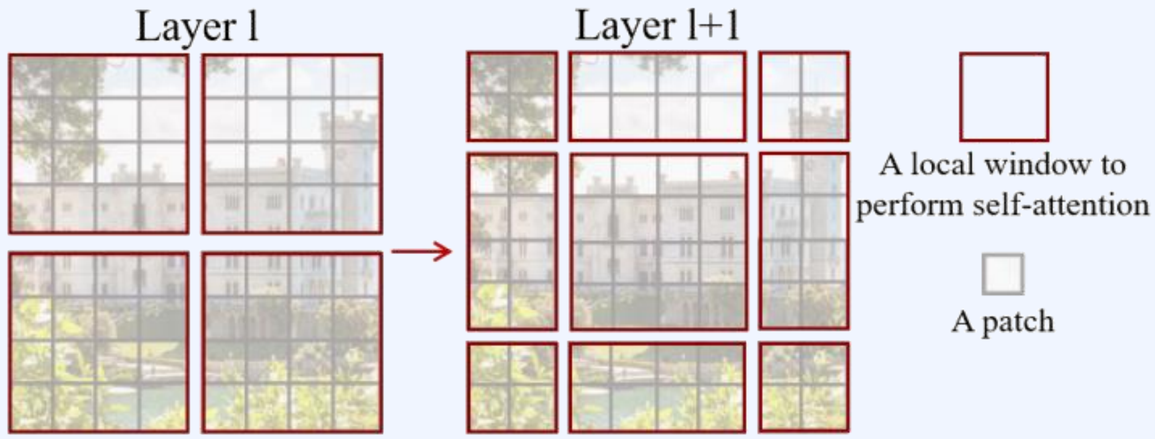
CeiT



Context Understanding
Visual Transformer

Z. Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. ICCV

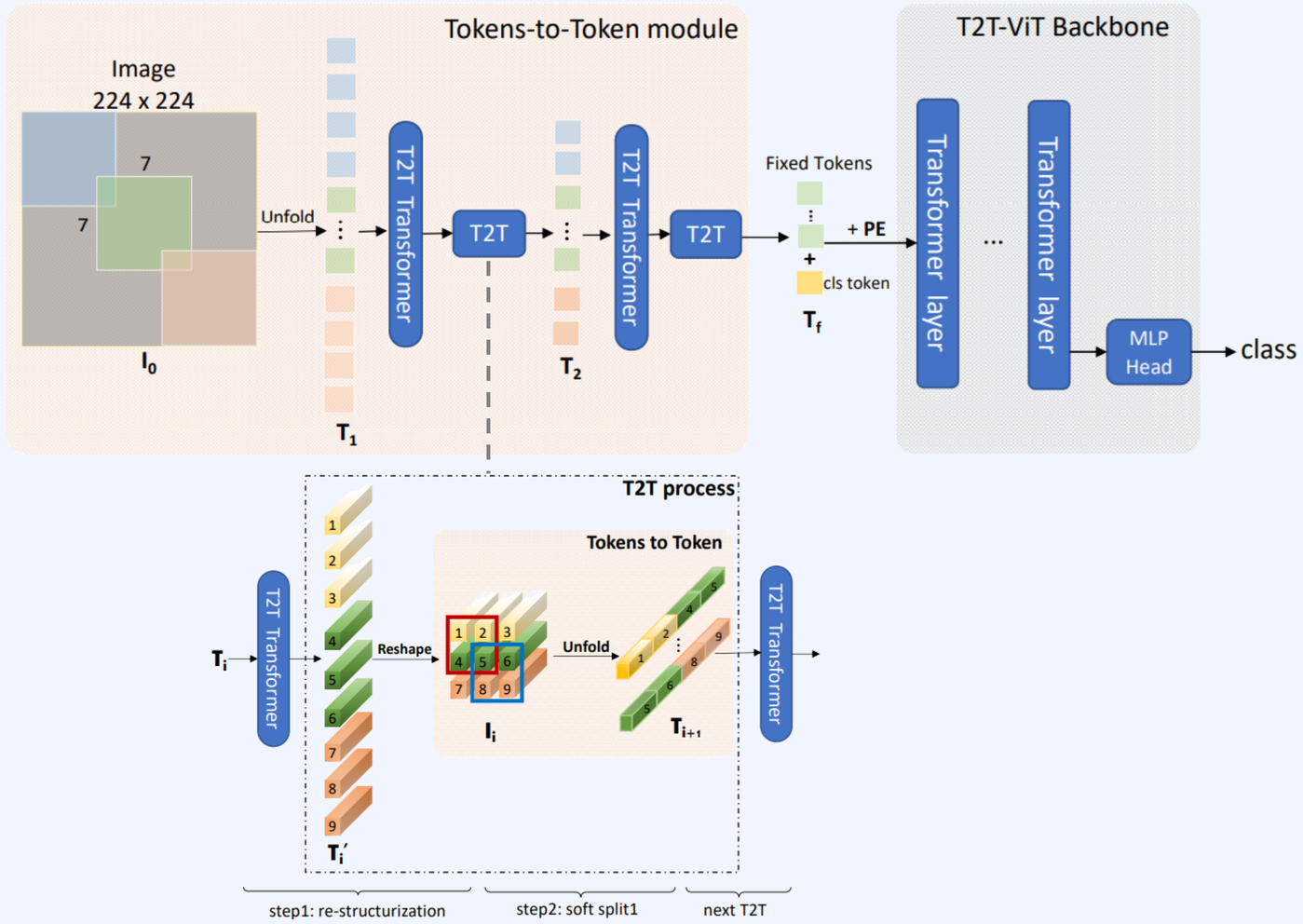




Context Understanding
Visual Transformer

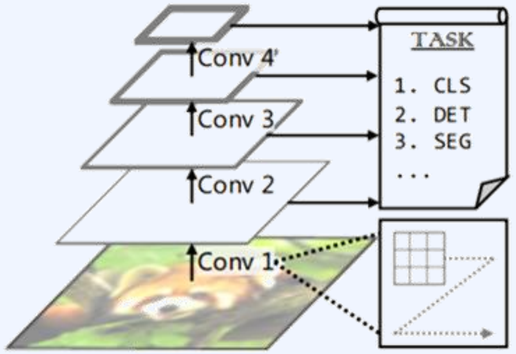
L. Yuan et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet. ICCV

T2T-ViT

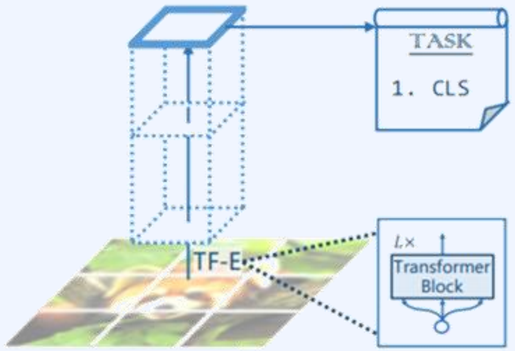


W. Wang et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. ICCV

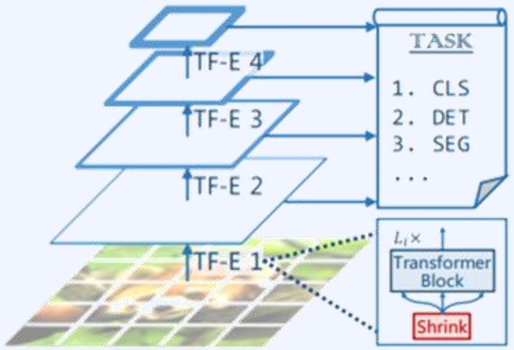
Pyramid Vision Transformer (PVT)



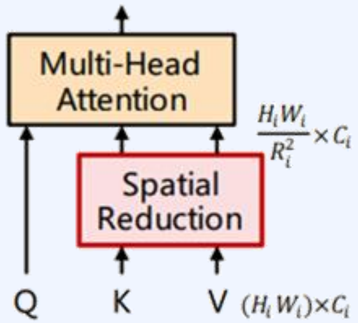
CNNs



ViT



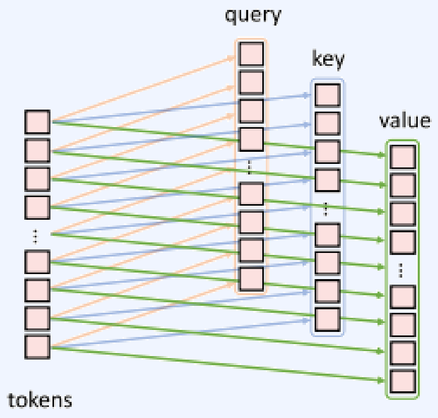
Pyramid Vision Transformer (PVT)



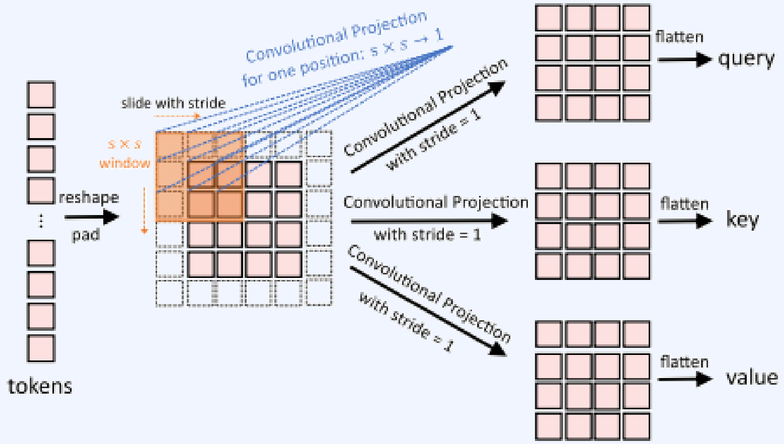
Spatial-Reduction Attention

Context Understanding
Visual Transformer

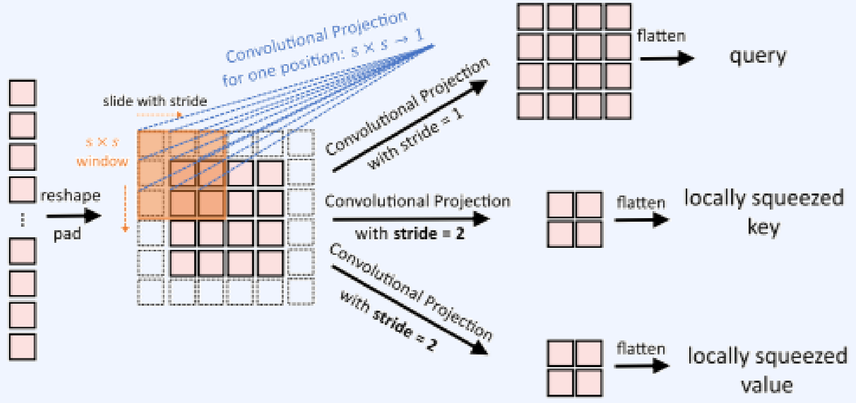
H. Wu et al. Cvt: Introducing convolutions to vision transformers. ICCV



Linear projection (ViT)



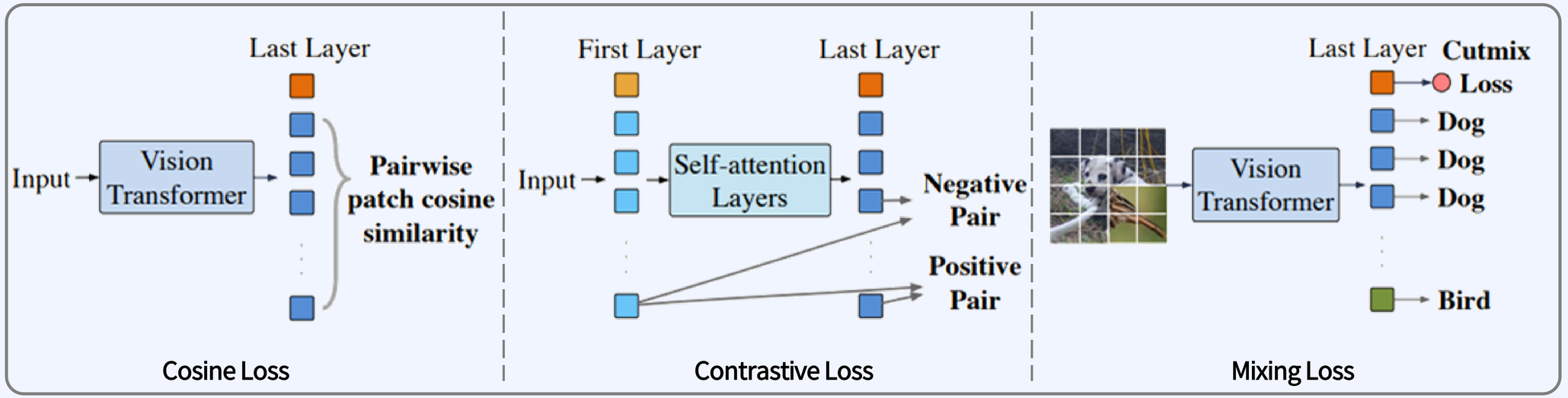
Convolutional projection



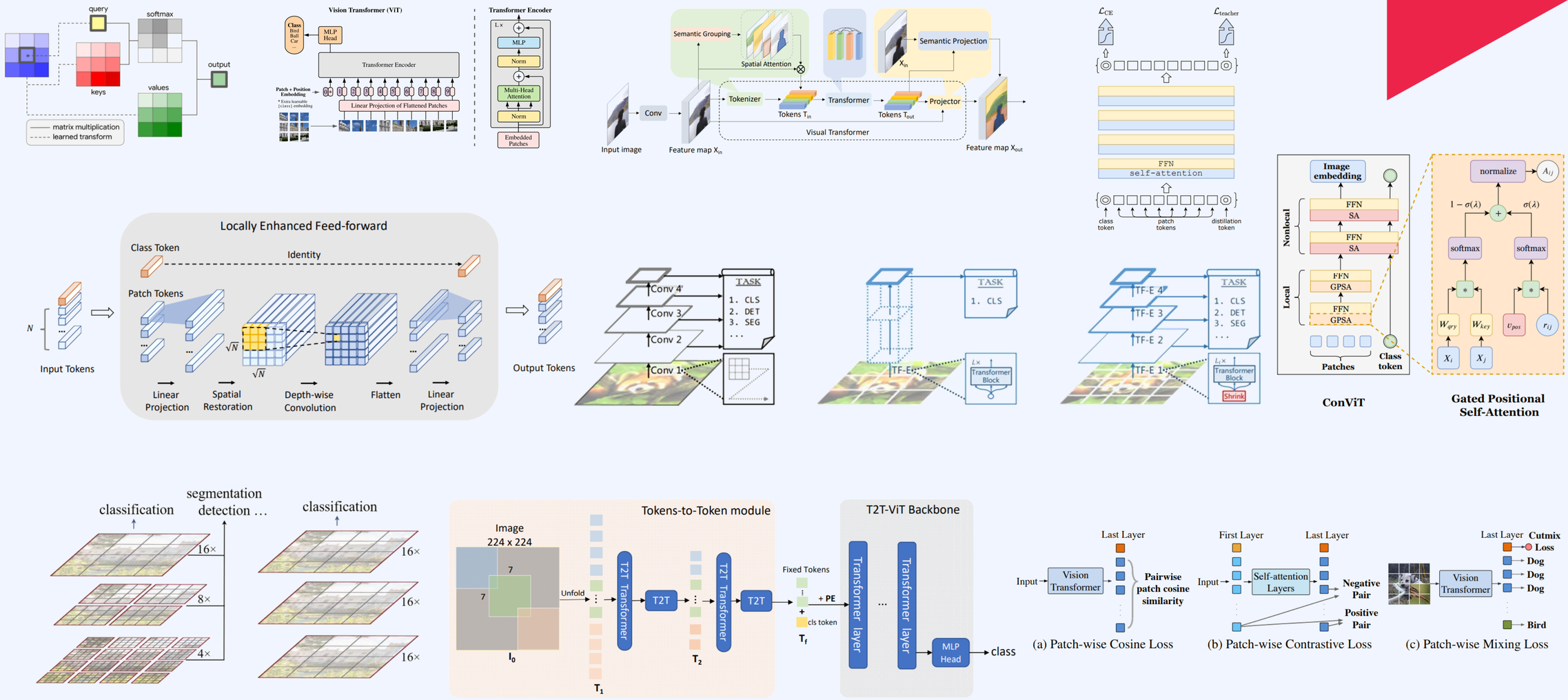
Squeezed convolutional projection (CvT)

Context Understanding
Visual Transformer

C. Gong et al. Vision transformers with patch diversification. arXiv:2104.12753



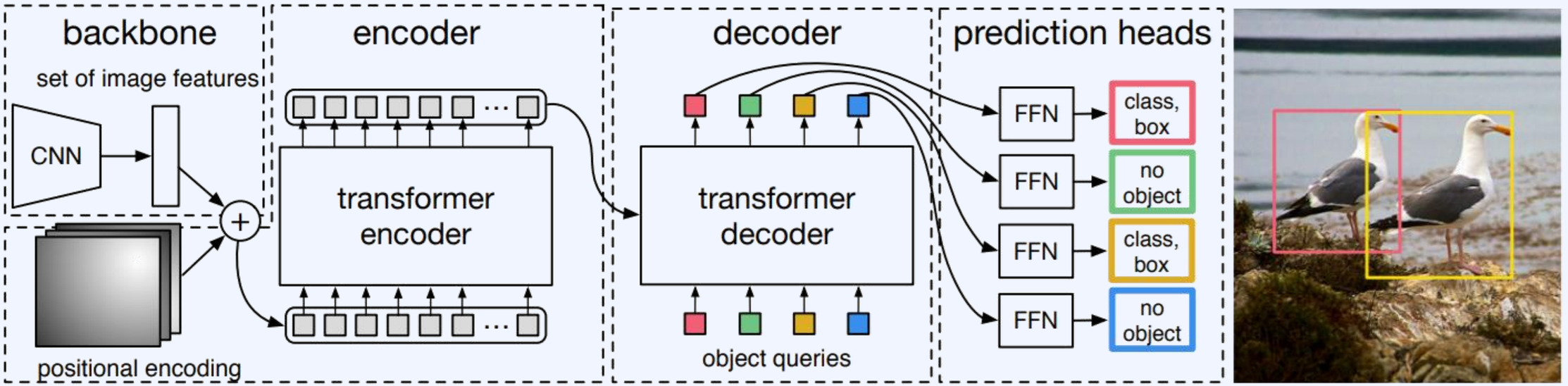
Context Understanding
Visual Transformer



Context Understanding
Visual Transformer

N. Carion et al. End-to-end object detection with transformers. ECCV

DETR

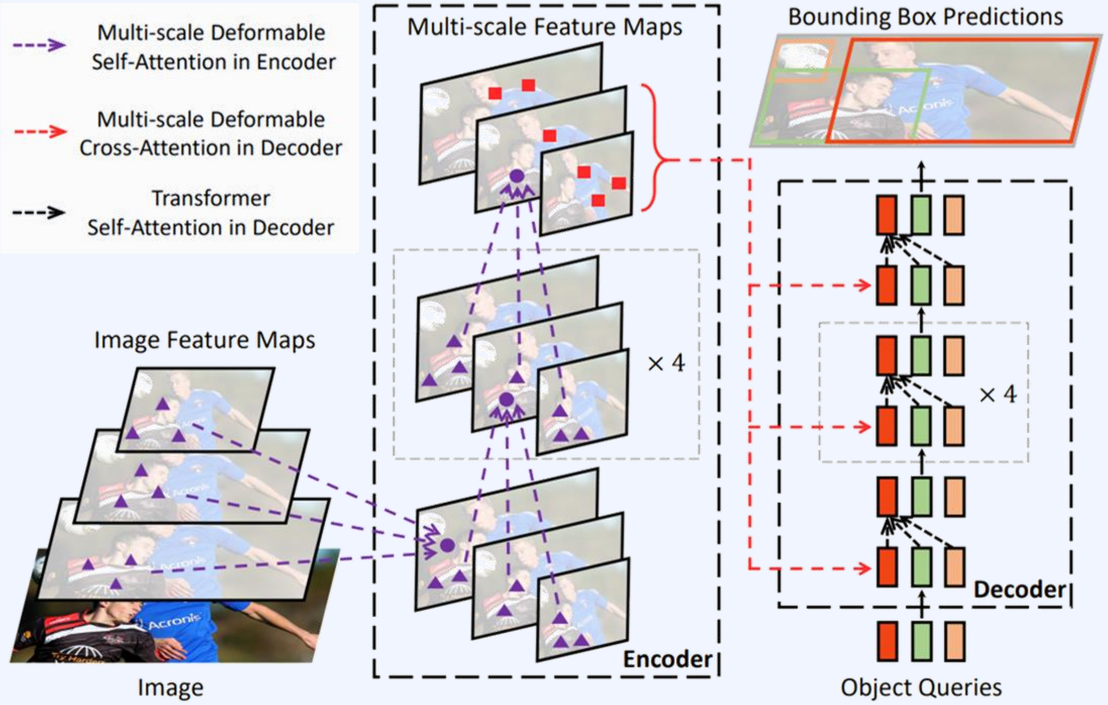


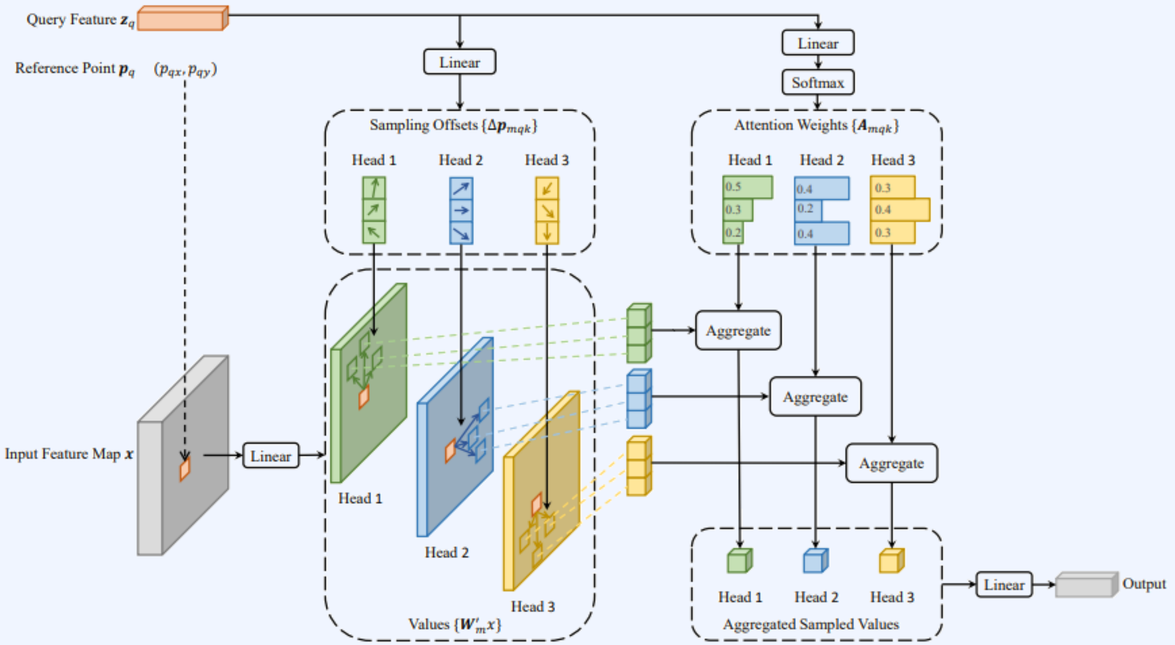
Deformable **DETR**, Sparse **DETR**, Conditional **DETR**, Anchor **DETR**, DAB-**DETR**, Efficient **DETR**, Dynamic **DETR**, DN-**DETR**, UP-**DETR**, FP-**DETR** , Panoptic **DETR**, Cell-**DETR**, **DETR**3D, M**DETR**, Tube**DETR**, ⋯

Context Understanding
Visual Transformer

X. Zhu et al. Deformable detr: Deformable transformers for end-to-end object detection. ICLR

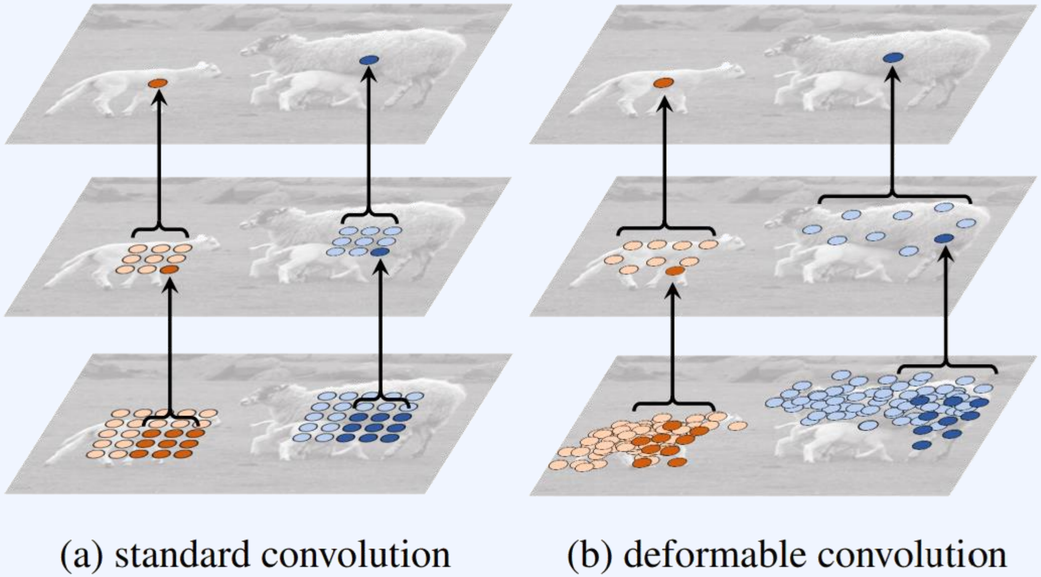
Deformable DETR



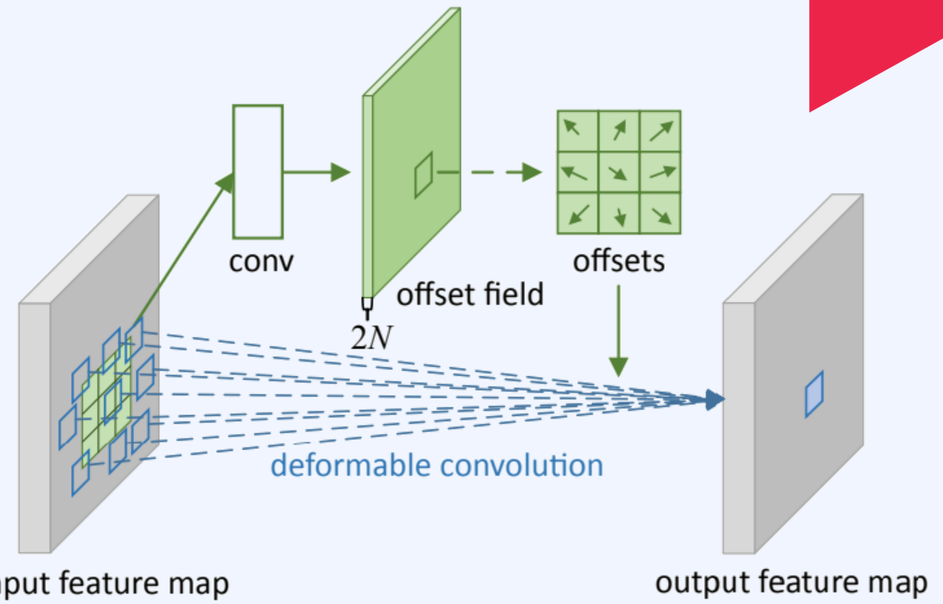


Context Understanding
Visual Transformer

J. Dai et al. Deformable convolutional networks. ICCV

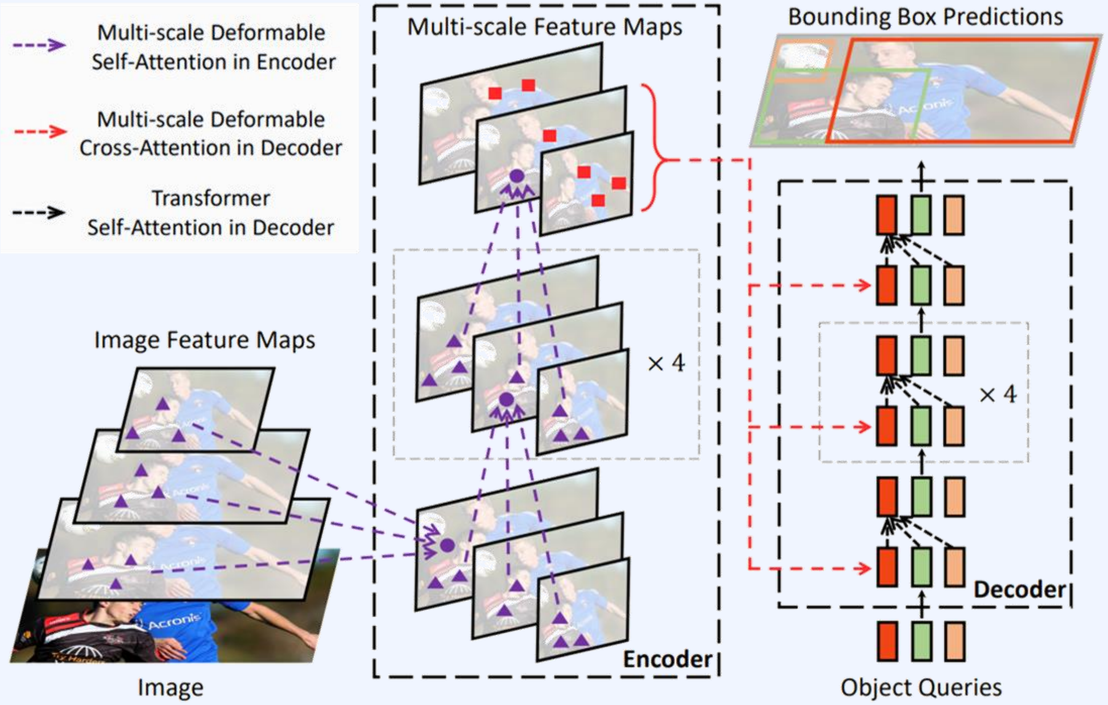


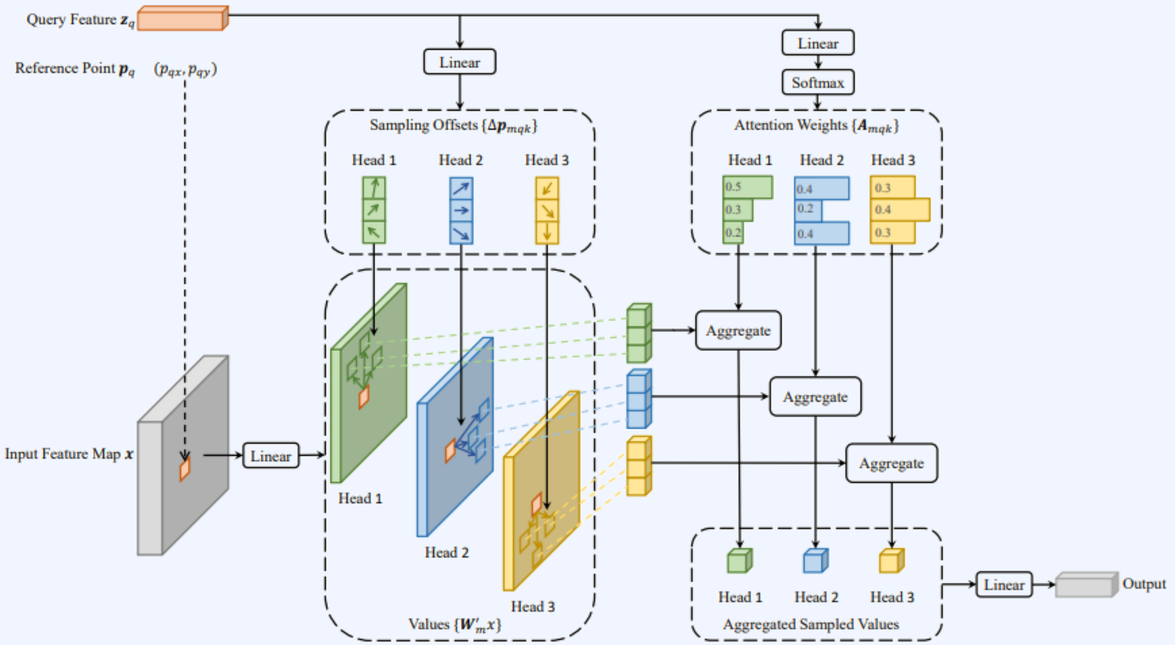
(a) standard convolution (b) deformable convolution



X. Zhu et al. Deformable detr: Deformable transformers for end-to-end object detection. ICLR

Deformable DETR

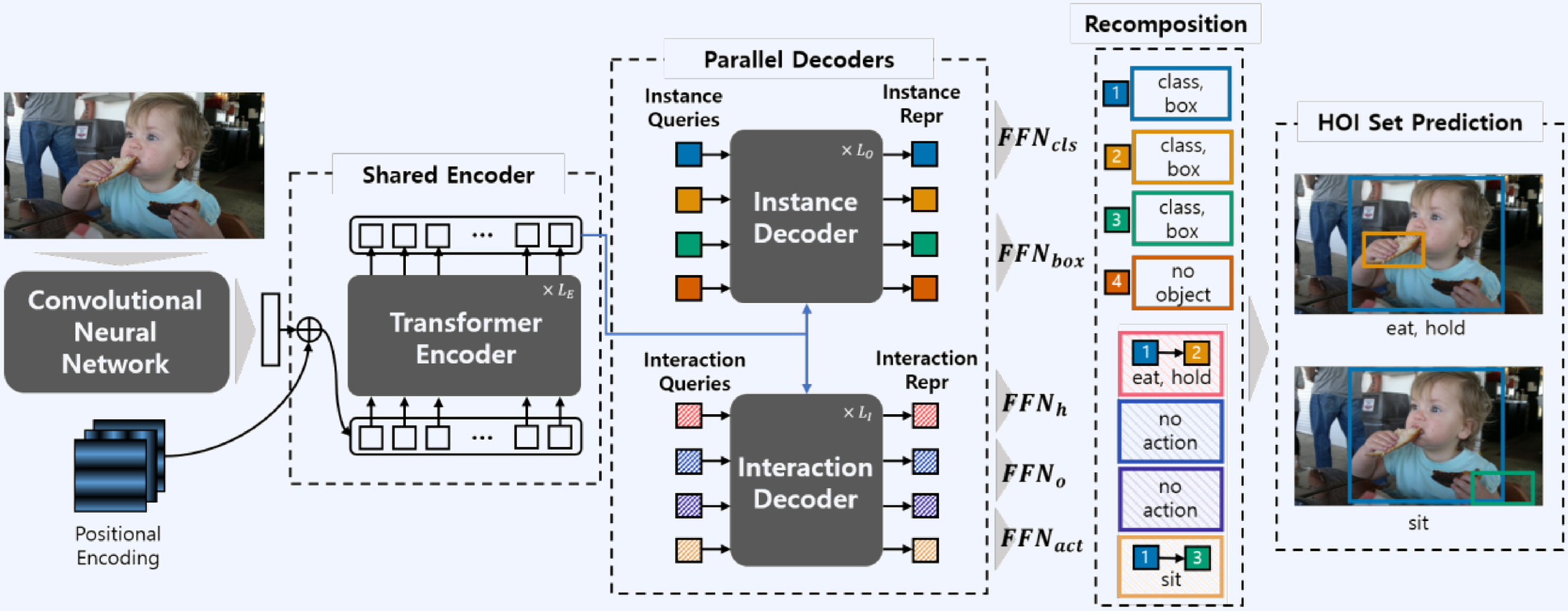




Context Understanding
Visual Transformer

B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

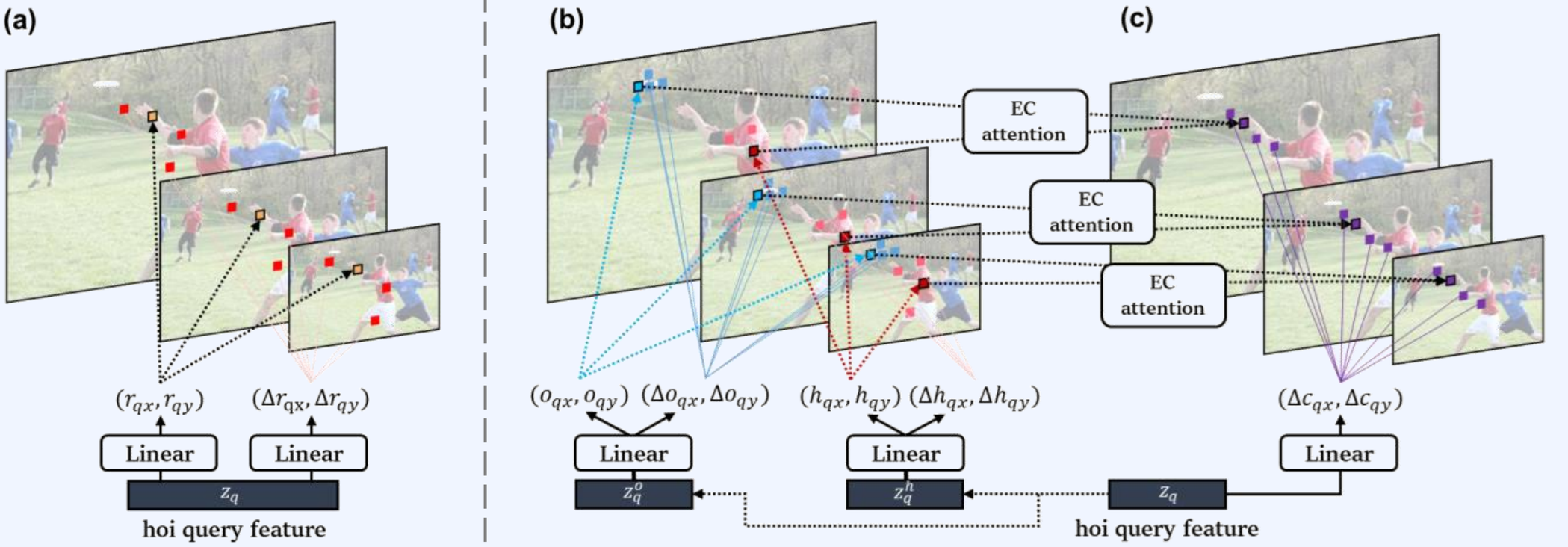
HOTR



Context Understanding
Visual Transformer

B. Kim et al. MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection. CVPR

MSTR

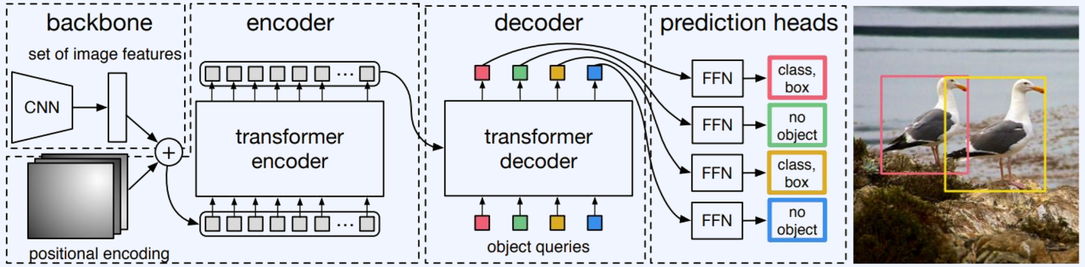


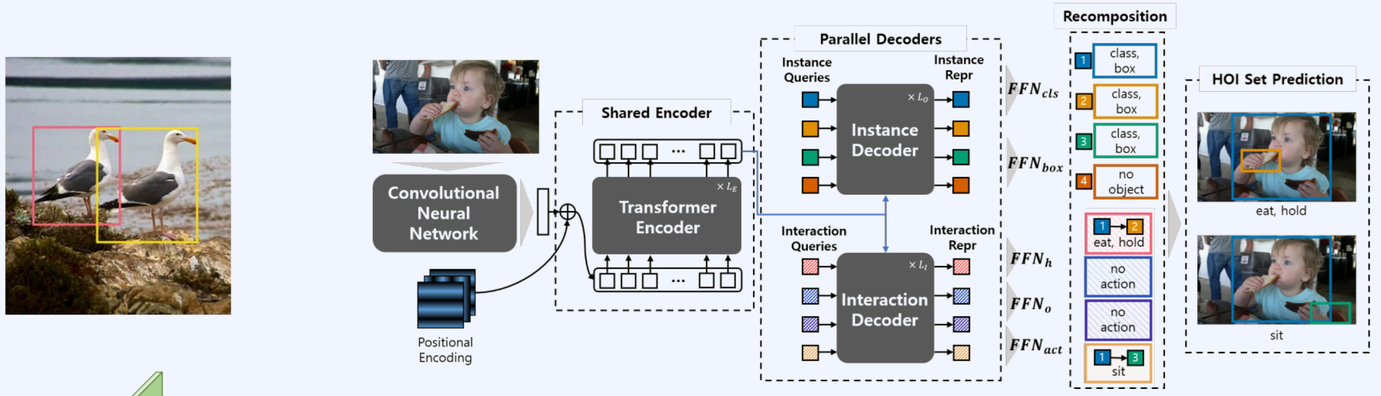
Deformable Attention

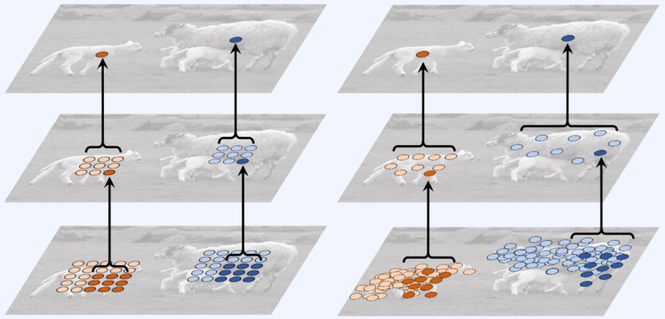
Dual-Entity Attention

Entity-conditioned Context Attention

Context Understanding
Visual Transformer







(a) standard convolution (b) deformable convolution