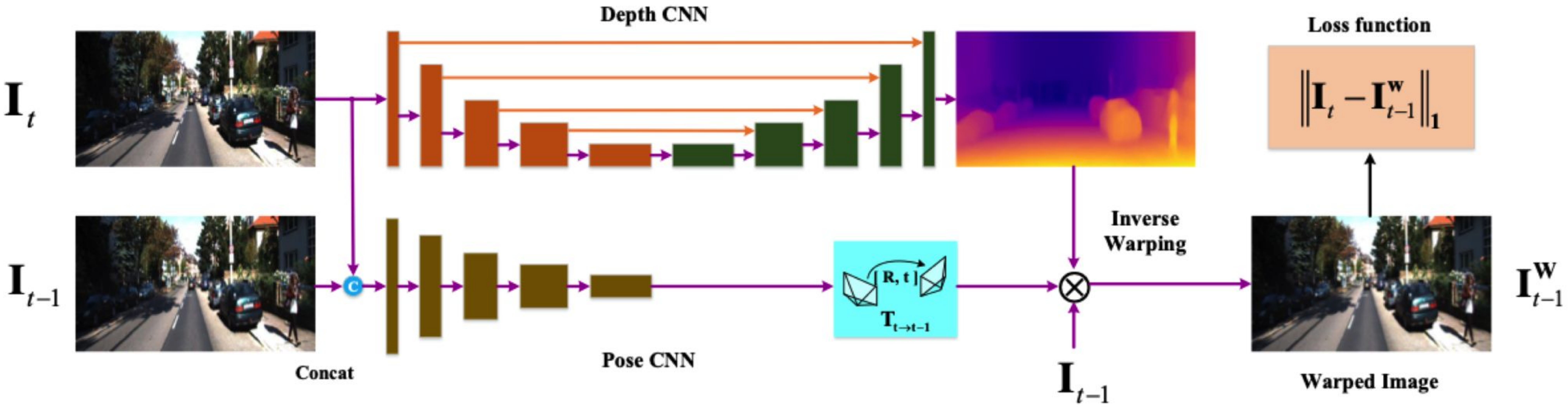
Self-Supervised depth estimation

Aleksei Karpov
Moscow Institute of Physics and Technology

Introduction

- Training:
  - Model for depth prediction using only geometrical constraints during training
  - Data for training: stereo pairs or image sequences
  - Target: estimate nearby frame
  - Loss function: photometric reprojection error
  - Model predict depth and egomotion between two frames (for image sequences)

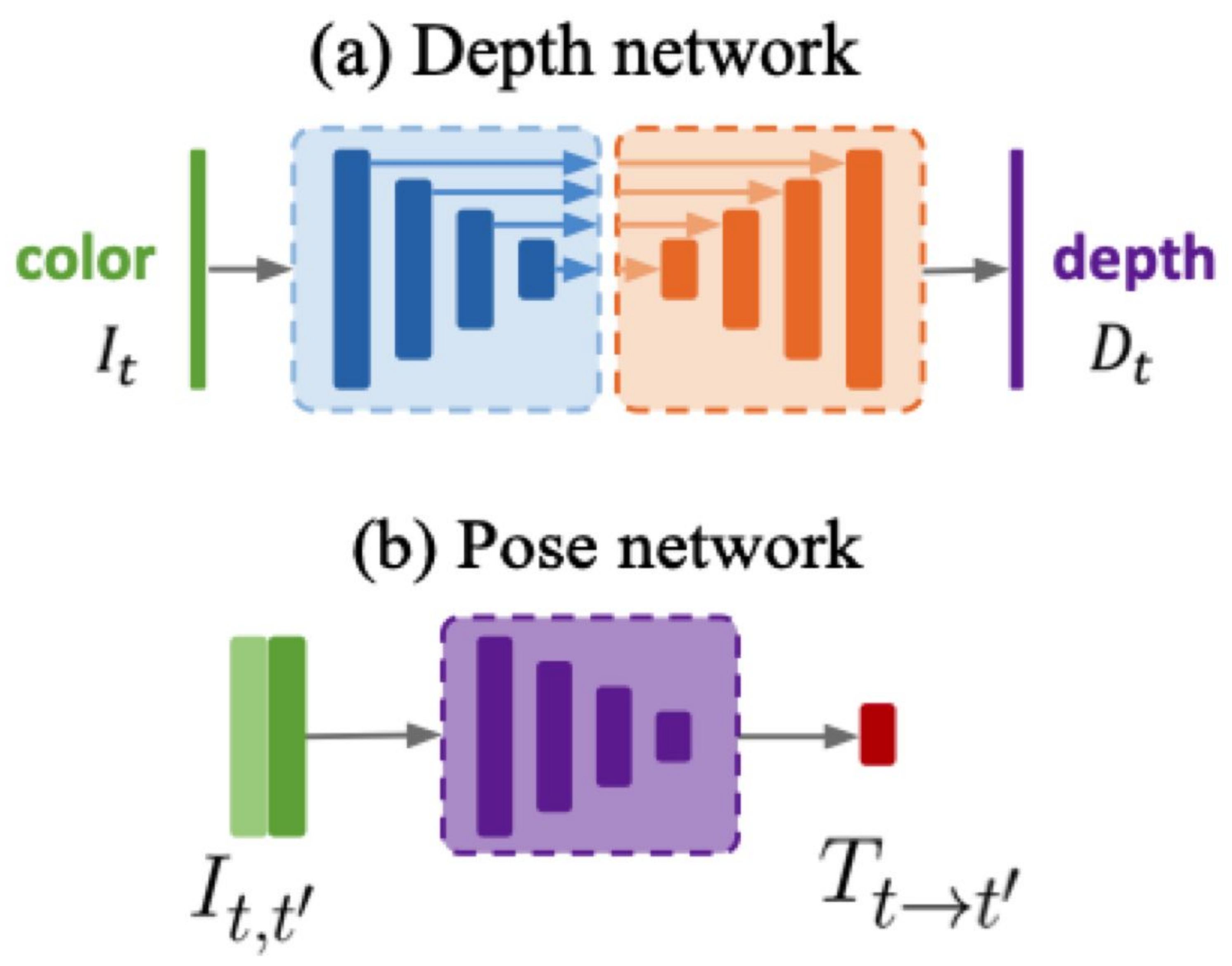
- Comparison with supervised
  - Easier data collection (not require LiDAR data)
  - Better generalization
  - Worse performance



Current architectures

- Fully convolutional networks for both networks
- U-Net based architectures for depth model:
  - Encoder: ResNet (Usually ResNet18 pretrained on ImageNet)
  - Decoder: Convolutional architecture
- ResNet for pose model
- Papers: monodepth2, ManyDepth

- CNN limitations:
  - Small receptive field
  - Poorly process global features



Novel architectures

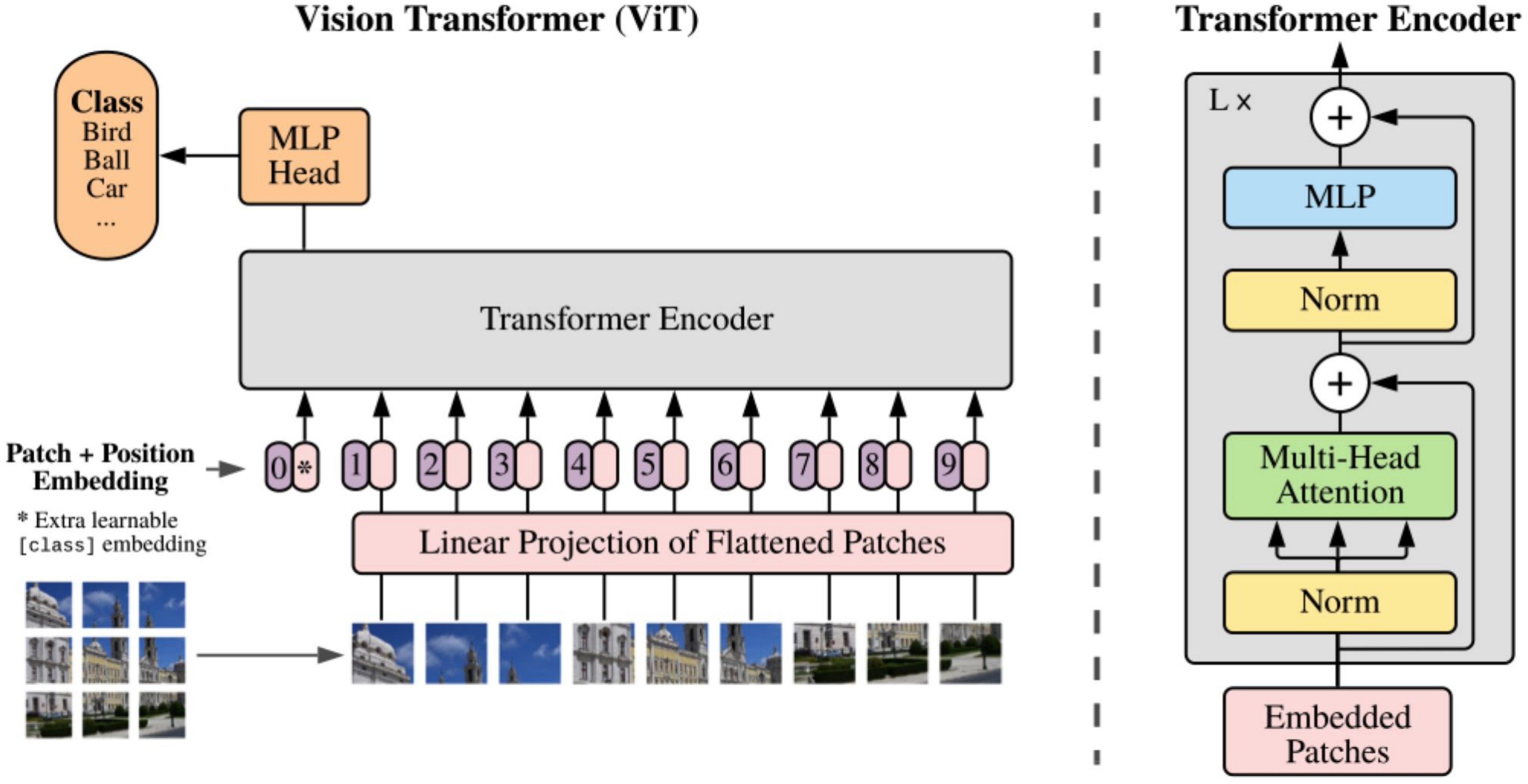
Vision Transformers (ViT)

- Split image into patches
- Embed patches
- Add position encoding
- Feed vectors to Transformer

- Advantages:
  - Global receptive field
  - Achieved promising results for classification and segmentation

- Disadvantages:
  - High complexity
  - Position encoding is fixed-dimension vectors

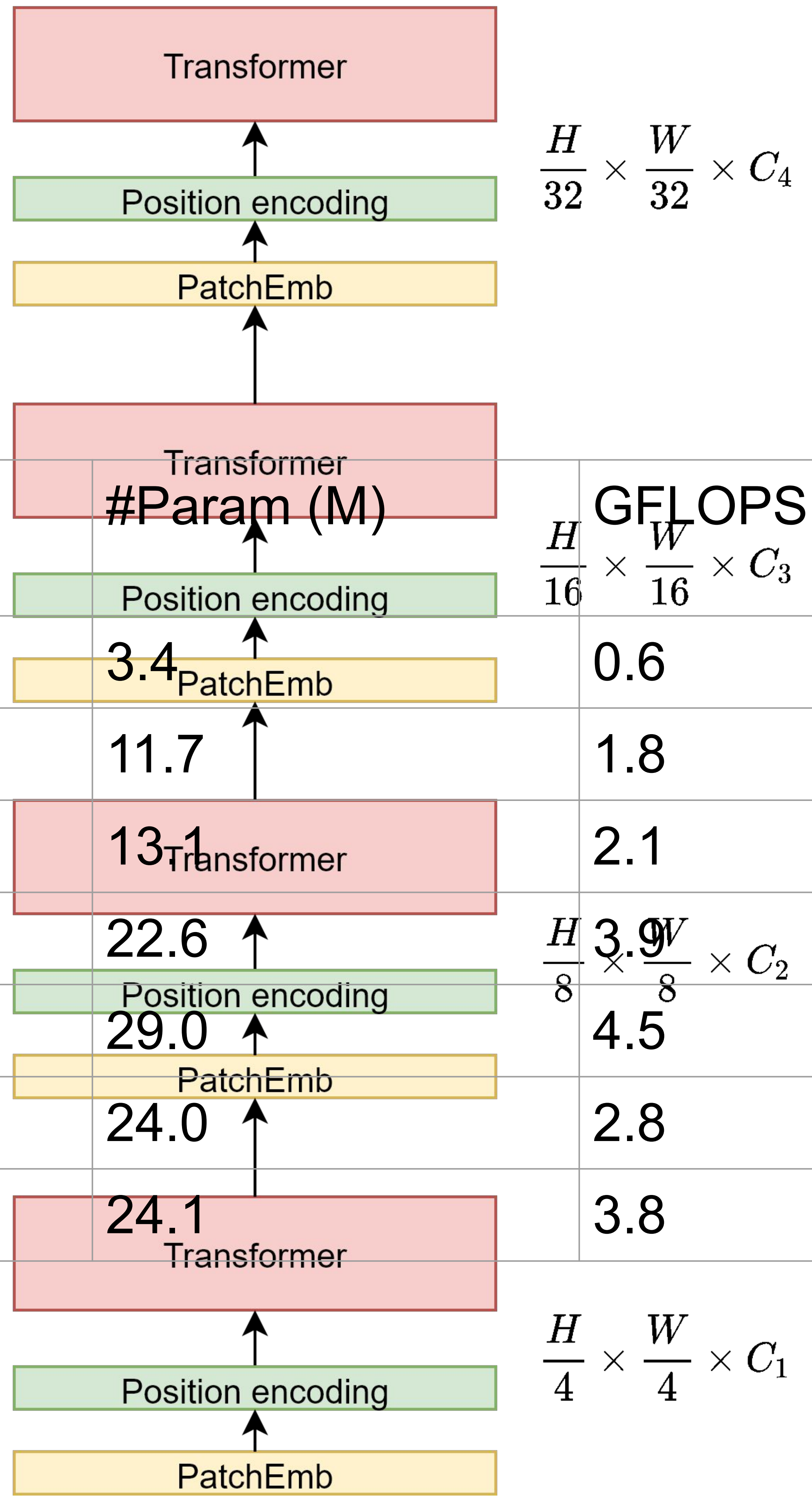
- [Original ViT paper]()
- [ViT for supervised depth prediction]()



ViT improvements

- Progressive shrinking to reduce complexity
- Overlapped patches
- Conditional position encoding
- Spatially sparable self-attention
- Papers: PVTv1, PVTv2, Swin, Twins

- Main improvements:
  - Less complexity(comparable with ResNet18)
  - Better performance
  - More suitable for dense prediction tasks

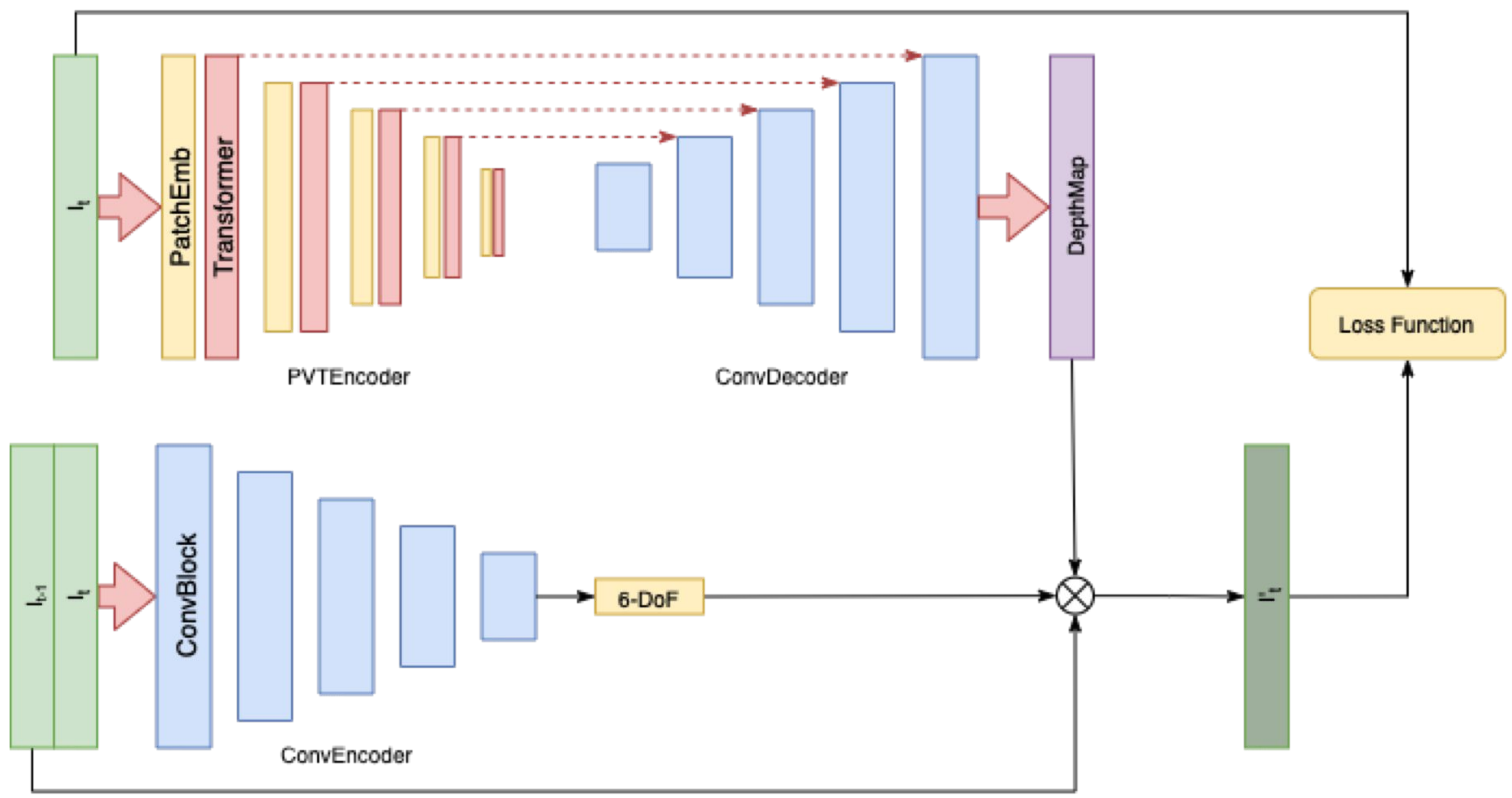

Transformer
$\frac{H}{32} \times \frac{W}{32} \times C_4$
Position encoding
PatchEmb
Transformer
$\frac{H}{16} \times \frac{W}{16} \times C_3$
Position encoding
PatchEmb
Transformer
$\frac{H}{8} \times \frac{W}{8} \times C_2$
Position encoding
PatchEmb
Transformer
$\frac{H}{4} \times \frac{W}{4} \times C_1$
Position encoding
PatchEmb

Encoder architecture	#Param (M)	GFLOPS
PVTv2_B0	3.4	0.6
ResNet18	11.7	1.8
PVTv2_B1	13.1	2.1
PVTv2_B2	22.6	3.9
Swin-T	29.0	4.5
Twins-SVT-S	24.0	2.8
Twins-PCPVT-S	24.1	3.8

ViT for self-supervised depth estimation

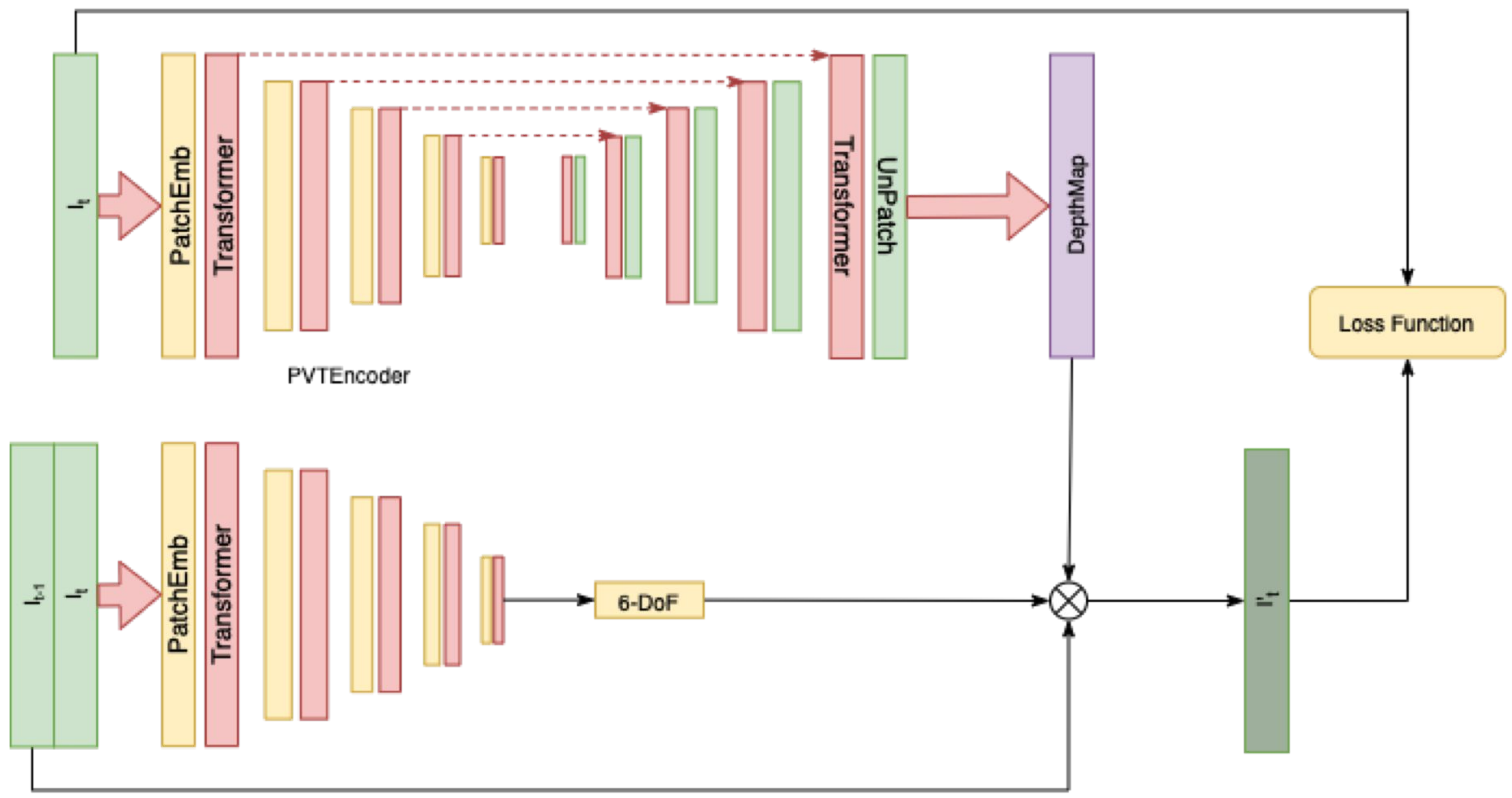
ViT Encoder

- Encoder based on PVT models
- Decoder contains CNN blocks
- Encoder weights pretrained on ImageNet
- Stable training process



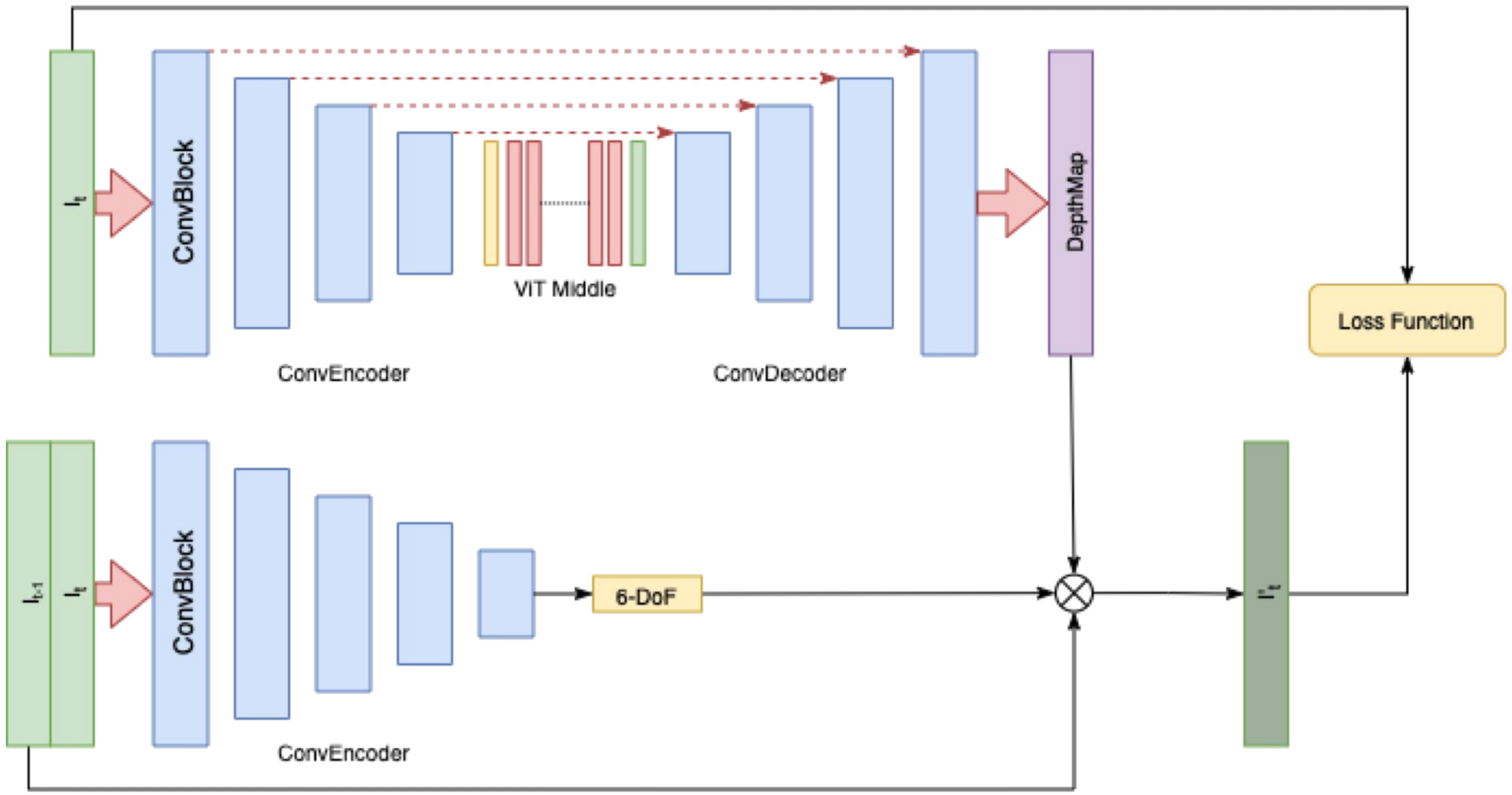
Pure ViT model

- Encoder based on PVT model
- Encoder contains ViT blocks
- Encoder weights pretrained on ImageNet
- Unstable training process



ViT Middle

- Encoder is ResNet18
- Middle based on ViT
- Decoder contains CNN blocks
- Encoder weights pretrained on ImageNet
- Stable training process



Conclusion

- **Current results:**
    - Overview current self-supervised depth estimation architectures
    - Analyze Vision Transformer architectures
    - Propose Transformer architecture for self-supervised depth estimation
    - Novel architecture better process global features

- **Future work:**
    - Train models with different ViT encoder architectures*
    - Train models with various CNN and ViT combinations*
    - Compare models performance
    - Implement cost volume based depth estimator like ManyDepth

* - partially done

Thank you for attention