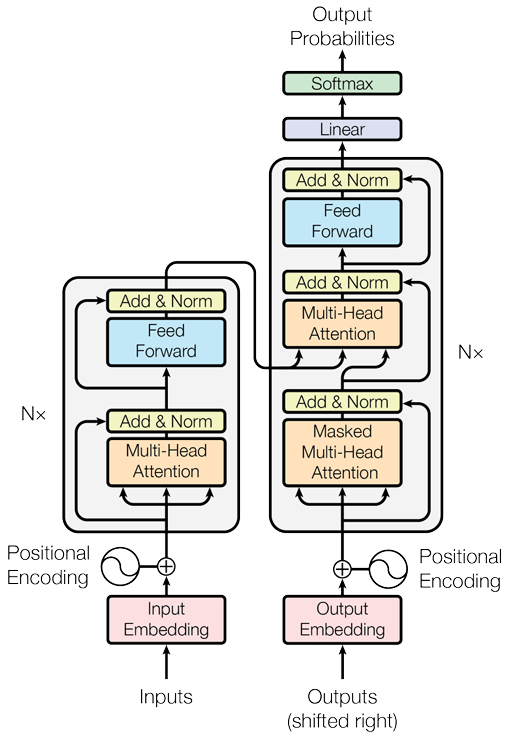
# Vision Transformers

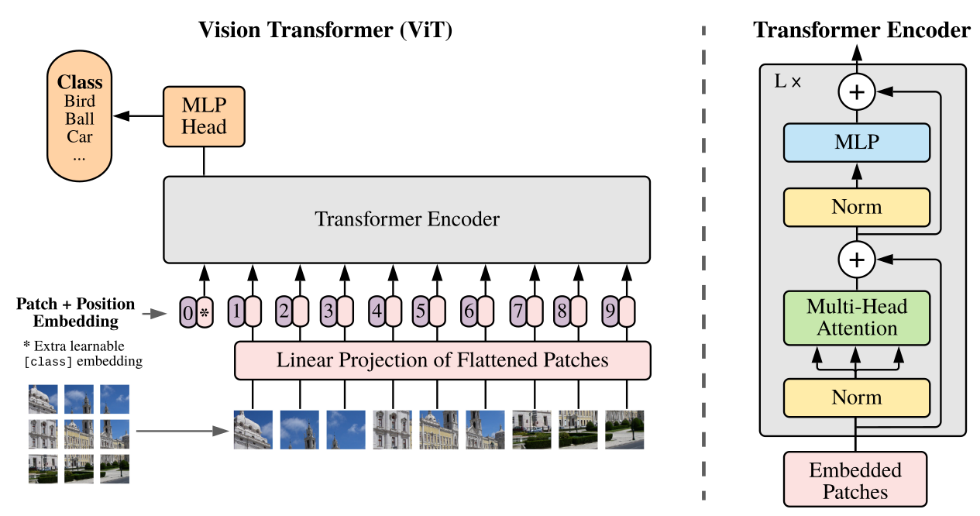
Transformers, originally introduced in NLP domain, have been successfully applied to Computer Vision Tasks. The original Transformer paper, consists of an encoder and decoder, containing several blocks of same architecture. Specifically, each block consists of following layers:

* Multi-head Attention
* Feedforward network
* Residual connection
* Layer Normalization



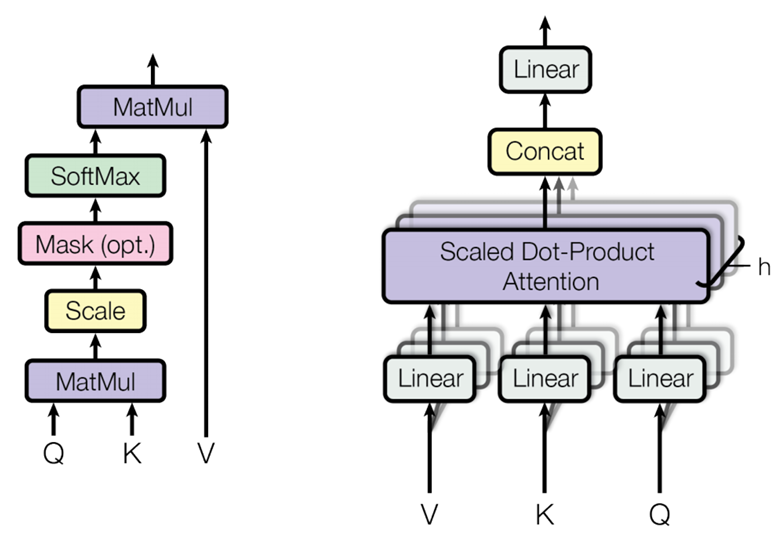
[Image reference](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

The architecture was adapted to computer vision tasks, originally in Vision Transformer (ViT). ViT uses only the encoder module of the NLP version. It treats patches of input images as tokens, computes the global relationship b/w tokens to create image embeddings



[Image reference](https://arxiv.org/abs/2010.11929)

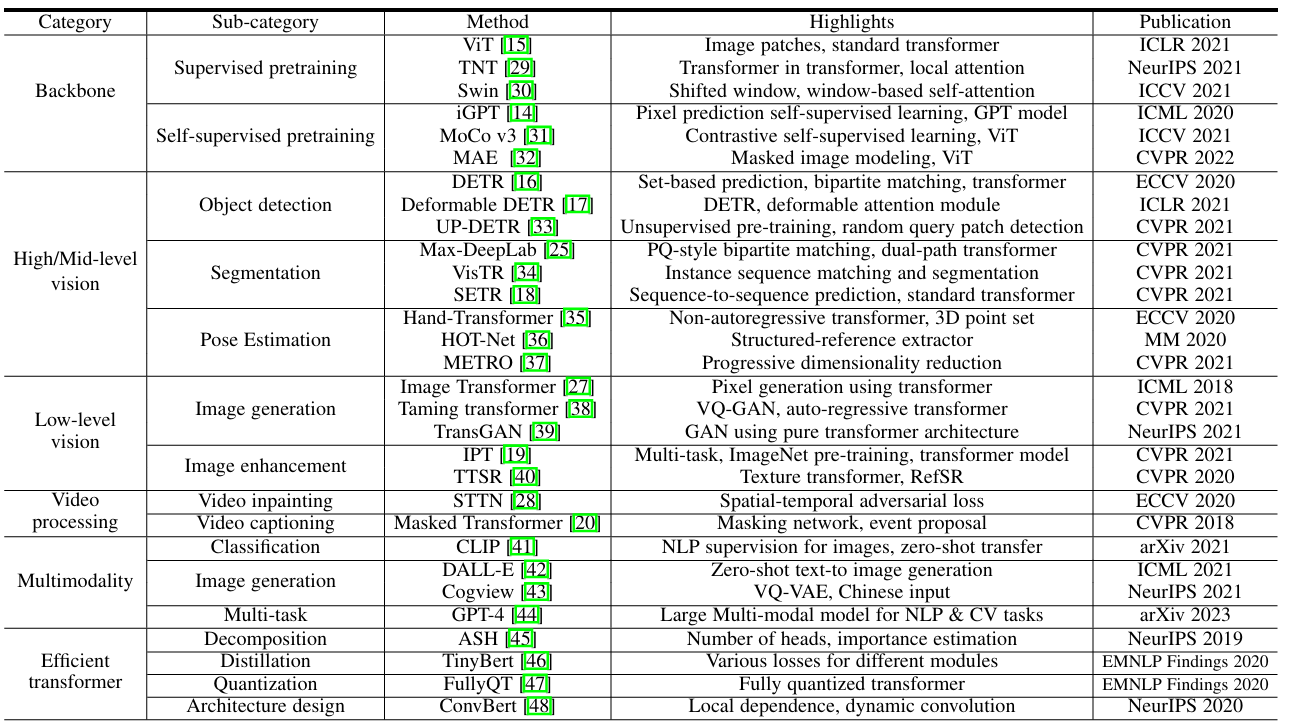
ViT used Multi-head Dot-product Self-attention mechanism in Transformer blocks as shown below:



[Image reference](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)

In such a self-attention mechanism, the importance of each location in feature map is computed w.r.t all locations. This ensures Global receptive field for ViT.

After success of ViT, different variants of transformer have been successfully applied to different Computer vision tasks, which is summarised in following table:



[Image reference](https://www.arxiv.org/pdf/2012.12556)

* Backbone: Used as primary feature extractor
* High / Mid-level vision: Used to find relationships and extract useful information from features, usually extracted from CNN modules
  + Object Detection
  + Segmentation
  + Pose Estimation
  + Lane Detection
* Low-Level vision:
  + Super resolution
  + Image denoising
  + Neural style transfer
* Multi-modality
  + NLP, CV combined supervision
* Efficient Transformers
  + Knowledge Distillation, Quantization, Neural Architecture Search (NAS)

## Advantages of ViT

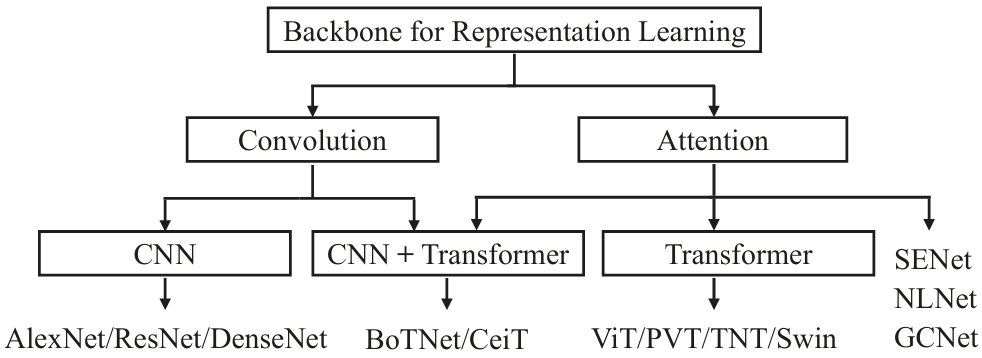
* Transformers can capture long distance characteristics, which derive global information
* Self-attention and Feed forward layers can be executed in parallel, making it more efficient and more suitable for Hardware accelerators, than RNNs, which predict sequentially
* Ability to handle sequential data makes it more suitable for handling temporal information, such as Tracking, Video Action recognition, captioning etc

## Disadvantages of ViT

* Self-Attention mechanism is compute-intensive and computes lot of memory, which makes it difficult to adapt to real-time applications and embedded devices

## Backbone for feature extraction

Transformers have been applied for feature extraction, either completely or combined with CNNs in various degrees.



[Image reference](https://www.arxiv.org/pdf/2012.12556)

### Pure Transformer

* ViT yields modest results when trained on mid-sized datasets such as ImageNet, achieving accuracies of a few percentage points below Resnets of comparable size.
* Reason is transformers don’t learn translation invariance and locality, as naturally as CNNs
* Pretraining on large datasets significantly improves performance. For example, ViT, when pretrained on JFM-300M dataset, and fine tunes on ImageNet, gives 88.36% accuracy
* Data efficient Image Transformers (DeiT) proposes a convolution free version of ViT, which combined with strong data augmentation, achieves impressive results on ImageNet dataset without any pretraining.
* Following areas are explored, for using Pure Transformers as Backbone
  + Improving local feature extraction – Vision Transformers uses a simple Linear layer to project the input 2D patch
  + Inter-head interaction -
  + Architecture – Dense prediction tasks like Object Detection, semantic segmentation depends on multi-scale features, which require FPN style Pyramid structures, as is explored in PVT, HVT, Swin Transformer etc

### Transformers with convolution

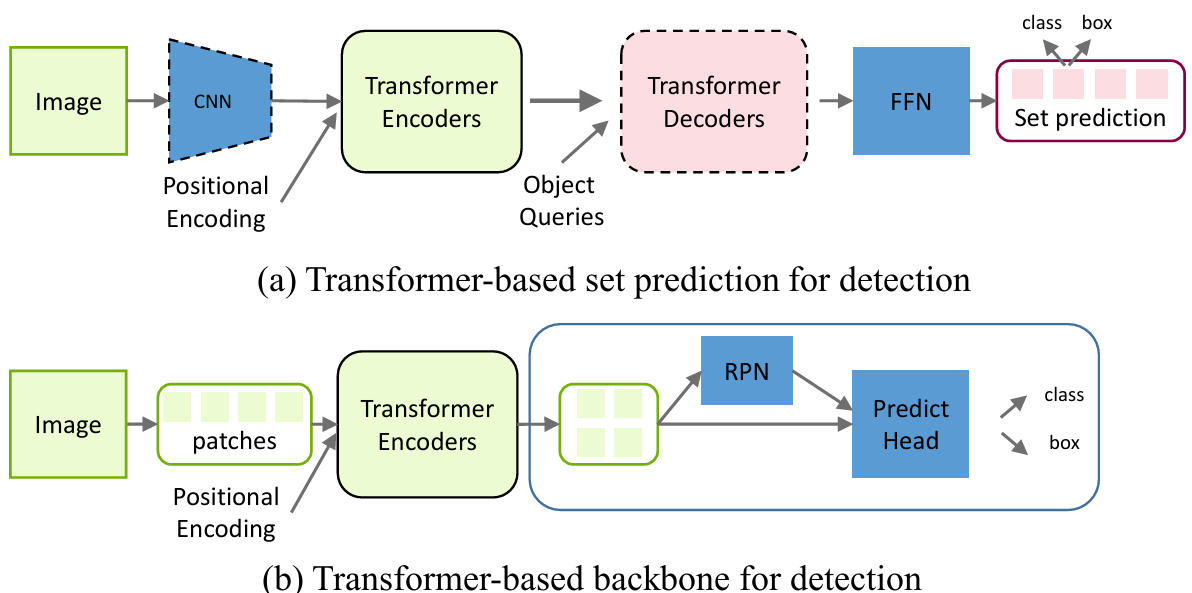
* CNNs are good at extracting local features, hence, when combined with transformers, improve performance considerably
* Transformers are also very sensitive to choice of optimizer, and other hyperparameters.
* When convolutional module is used in ViT, in the early preprocessing step, we’re able to train with AdamW, SGD optimizers, without significant drop in accuracy, while convergence is significantly improved

## High-level vision

### Object Detection

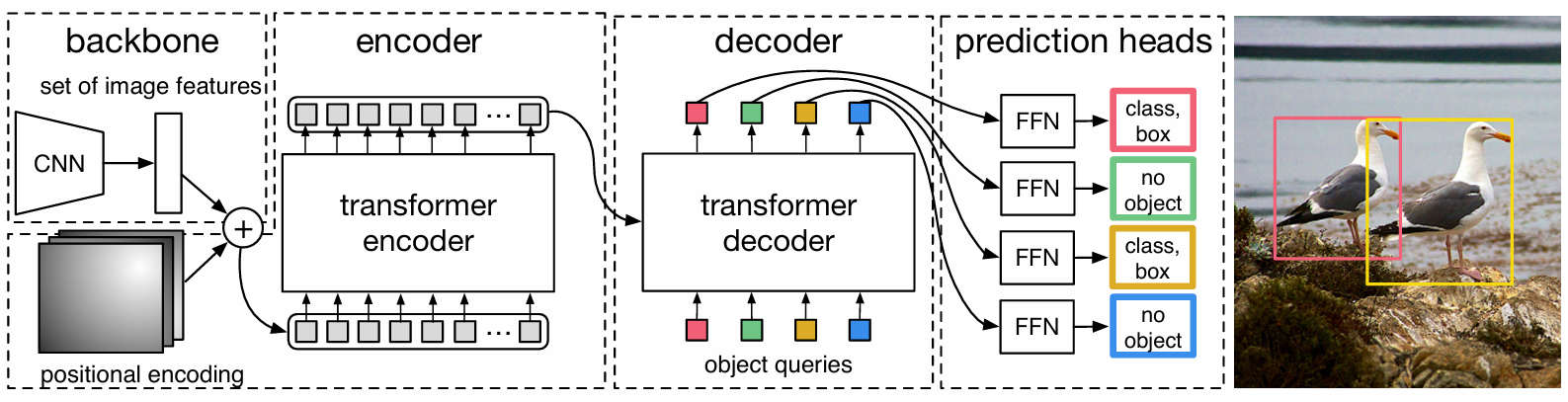
Traditional object detectors are mainly built upon CNNs, but transformer-based object detection has gained significant interest recently due to its advantageous capability. Transformer-based object detection methods are broadly categorized into following two groups:

* Transformer based set prediction methods (DETR, Deformable DETR)
* Transformer-based backbone methods (RetinaNet, Cascade RCNN)



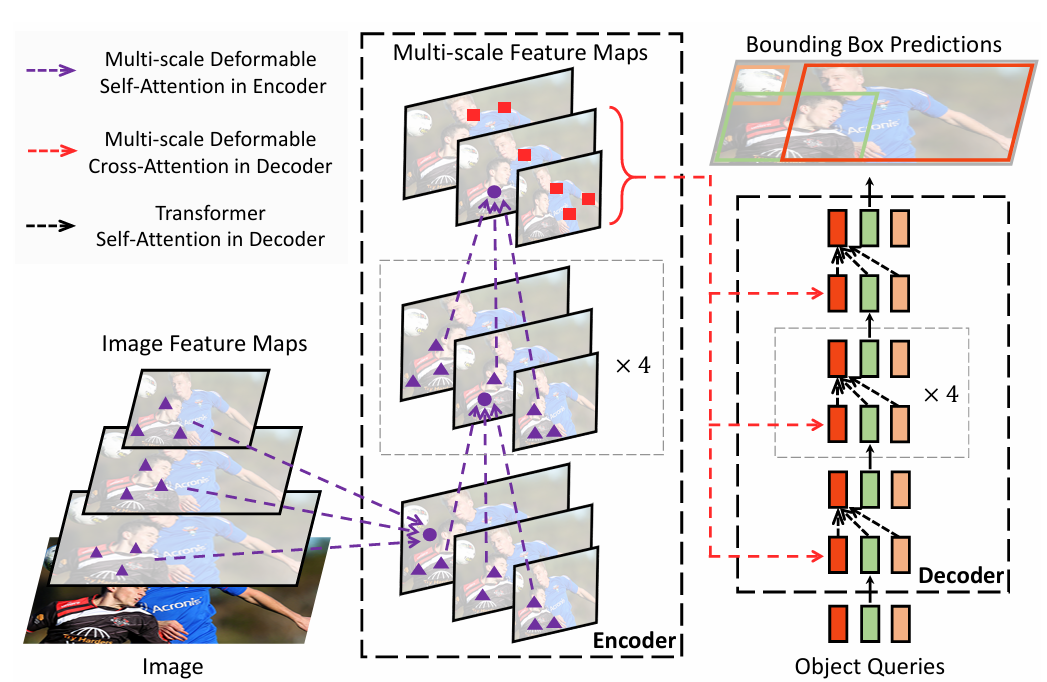
[Image reference](https://www.arxiv.org/pdf/2012.12556)

In DETR, a CNN backbone is used to extract features, after combining with position encoding, are used as input tokens to transformer encoder. The output from encoder is fed to decoder along with object queries. Finally, a MLP based prediction head, predicts the class and bounding box coordinates.



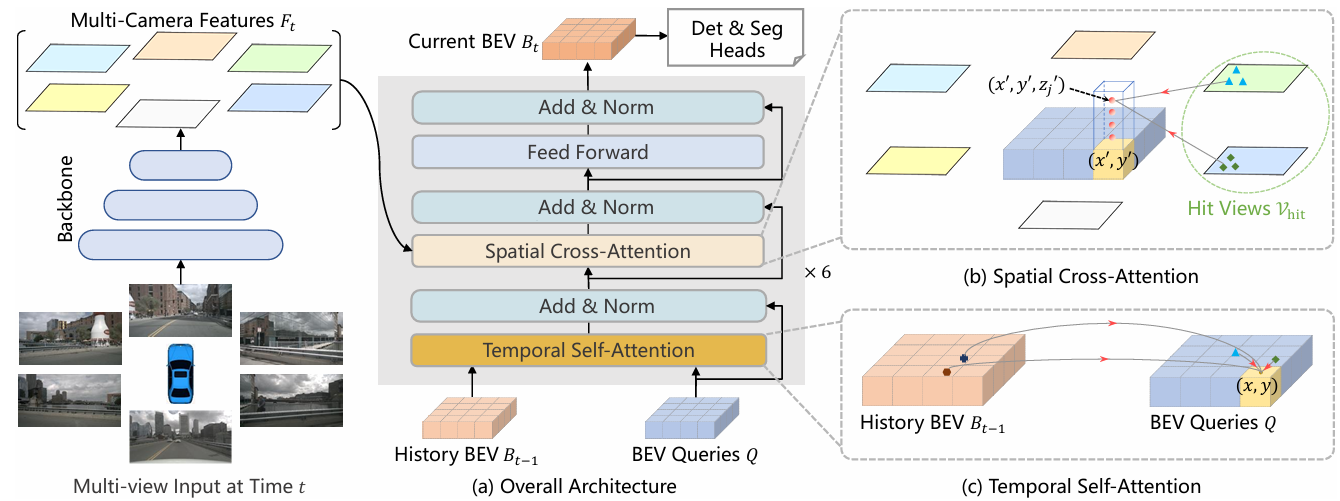
[Image reference](https://arxiv.org/pdf/2005.12872)

But vanilla DETR poses disadvantages such as longer training schedule and poor performance for small objects. Deformable DETR was introduced addresses this, by attending to a small set of key positions around a reference point rather than looking at all spatial locations on image feature maps. This reduces complexity and improves convergence, while enabling multi-scale fusion.



[Image reference](https://arxiv.org/abs/2010.04159)

Deformable attention is basic building block for various 3D dense prediction architectures such as BEVFormer, BEVFormer2.



[Image reference](https://arxiv.org/pdf/2203.17270)

In BEVFormer, we have multi-camera inputs at time t. Each of the camera image is passed through a shared Backbone network to extract baseline features. BEV query is constructed, based on the 3D volume of interest, based on specific resolution (e.g. 20cm cubes covering 50m x 50m x 5m cuboid). This BEV query, along with BEV feature from previous timestep (t -1) is passed through a temporal self-attention block to understand the changes in environment w.r.t previous timestep and also to infer velocity, acceleration of dynamic objects. Each query in output BEV feature map is lifted, (~repeated) like a pillar in 3D. Instead of computing attention of all BEV query points w.r.t each of extracted camera features, the query points are sampled and weighted sum of sampled features is calculated as output of spatial cross attention module.