1. Vision-Language Models in the era of LLMs
2. **ViT: From classification to detection, segmentation, …**

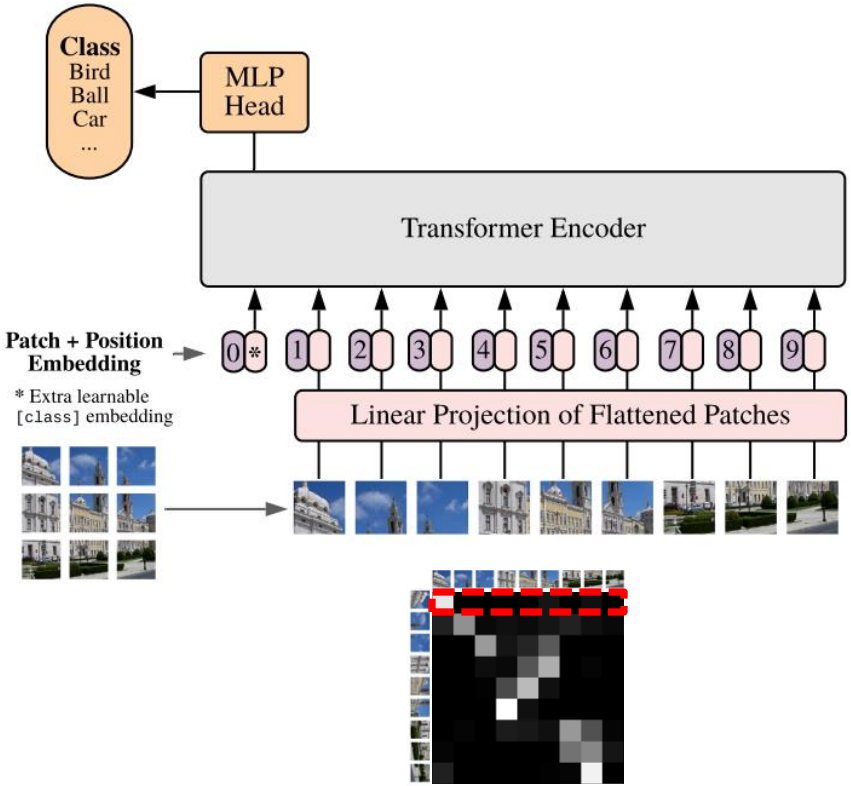
(Visual) Transformers

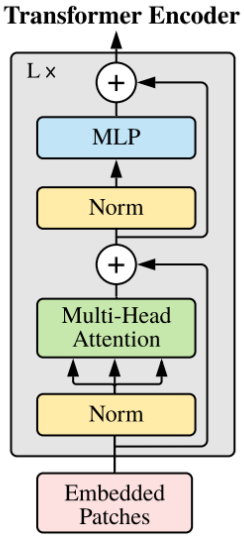
ViT (Vision Transformers) architecture
=> **Self attention encoder** modules for classification

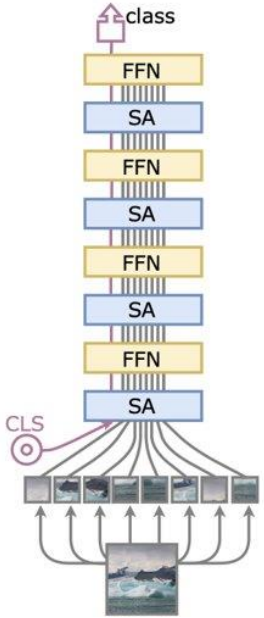
Published as a conference paper at ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†
*equal technical contribution, †equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com







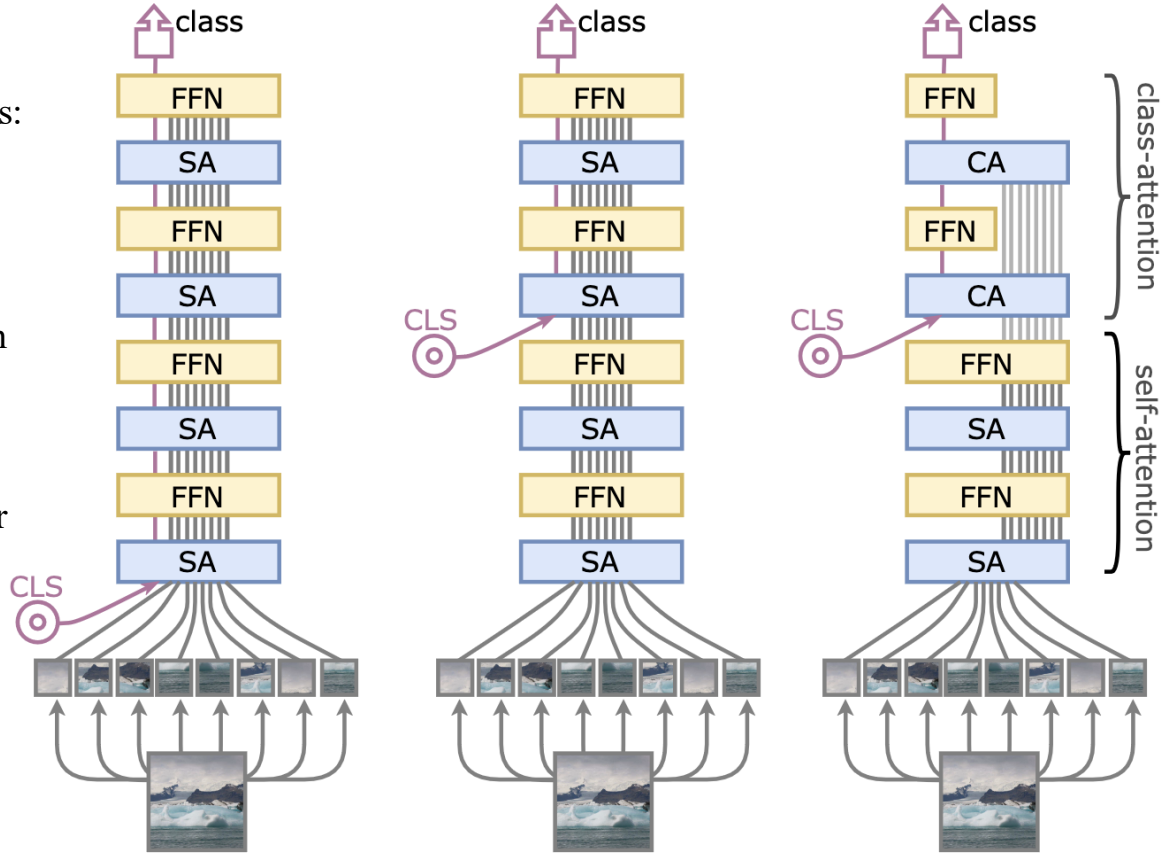
Class Activation architecture

In ViT class embedding CLS token inserted along with the patch embeddings:

- helping the attention process
- preparing the vector to be fed to the classifier

CaiT freezes the patch embeddings when inserting CLS:

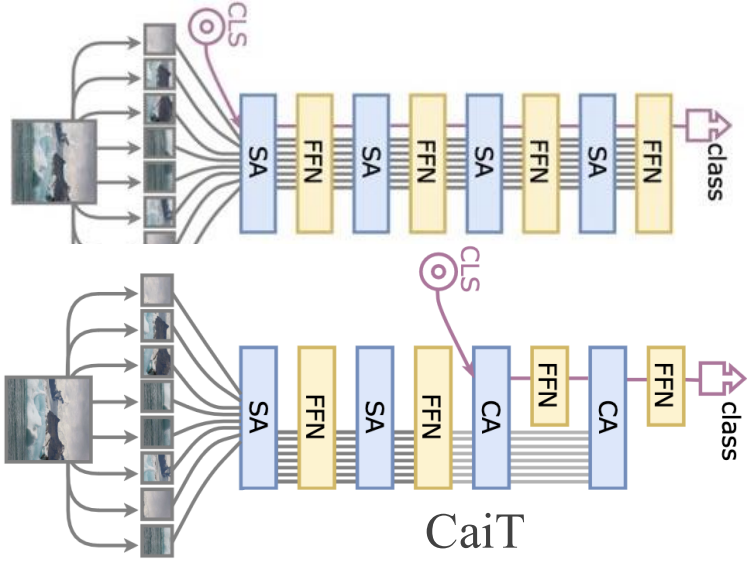
- last part of the network (2 layers) dedicated to summarizing the information to be fed to the classifier
- save compute



(Visual) Transformers for detection, segmentation, …

Design output for classification, detection, …

- CLS token for classification
- CaiT strategy: CLS to decode the embeddings

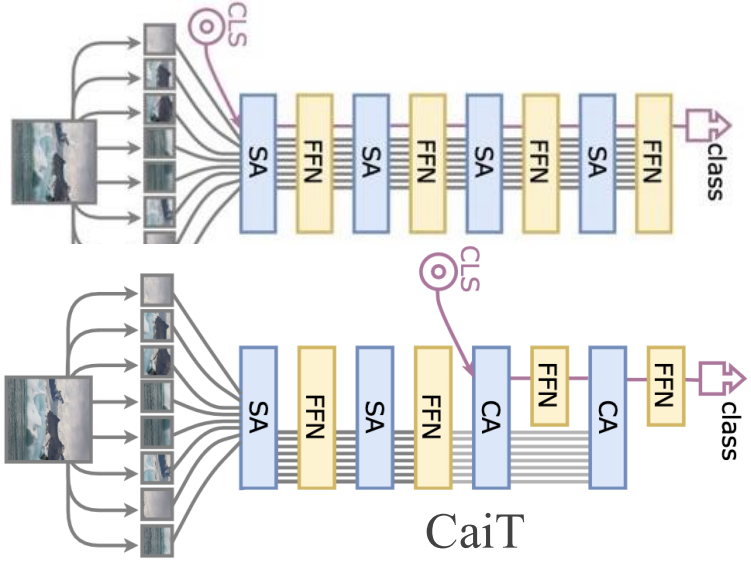


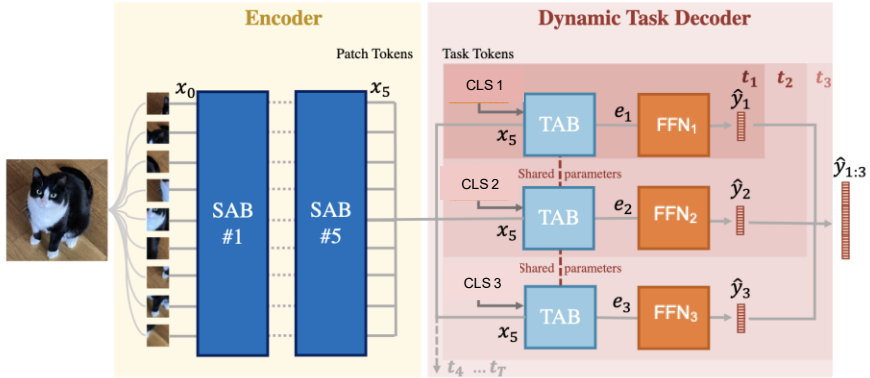
(Visual) Transformers for detection, segmentation, …

Design output for classification, detection, …

- CLS token for classification
- CaiT strategy: CLS to decode the embeddings
- Extension to incremental classification task learning:

And for other type of output as detection?

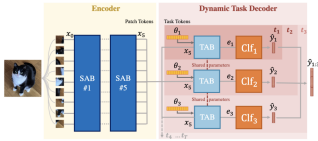


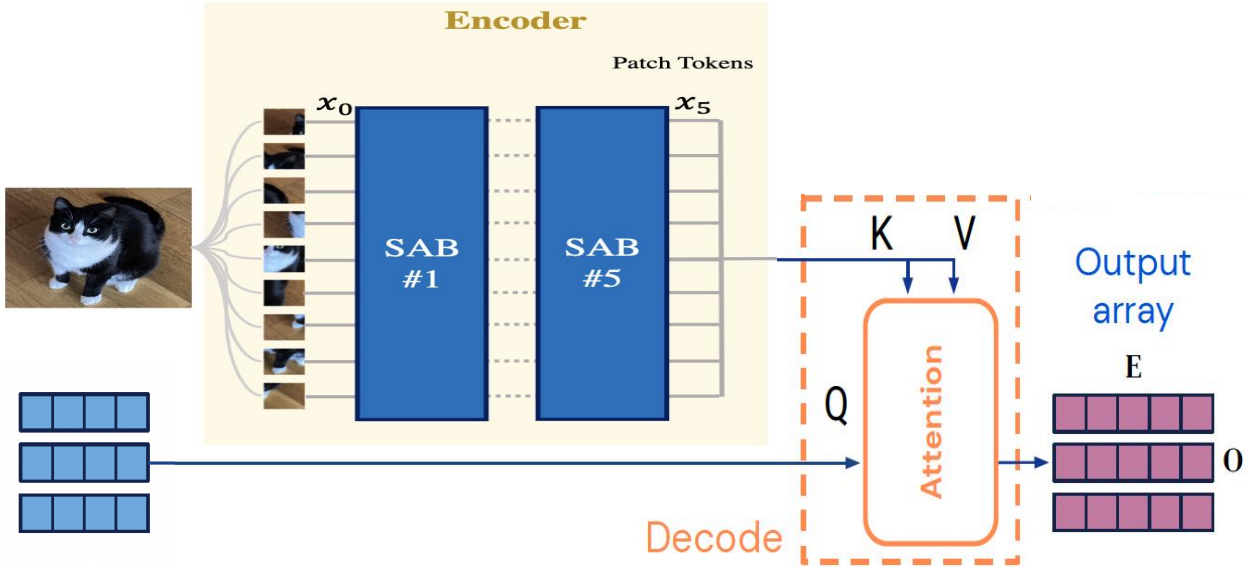


TAB: Task Attention Block

(Visual) Transformers for detection, segmentation, …

To summarize:



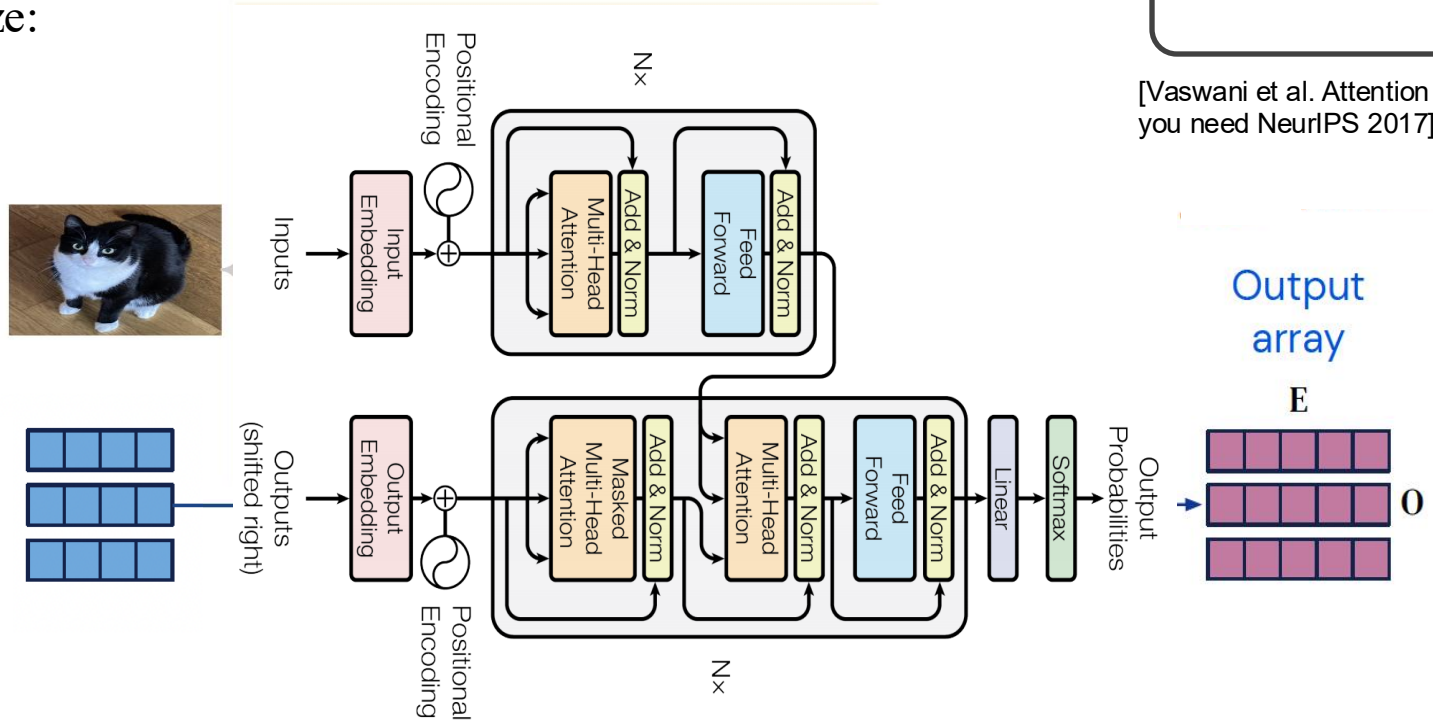


(Visual) Transformers for detection, segmentation, …

To summarize:

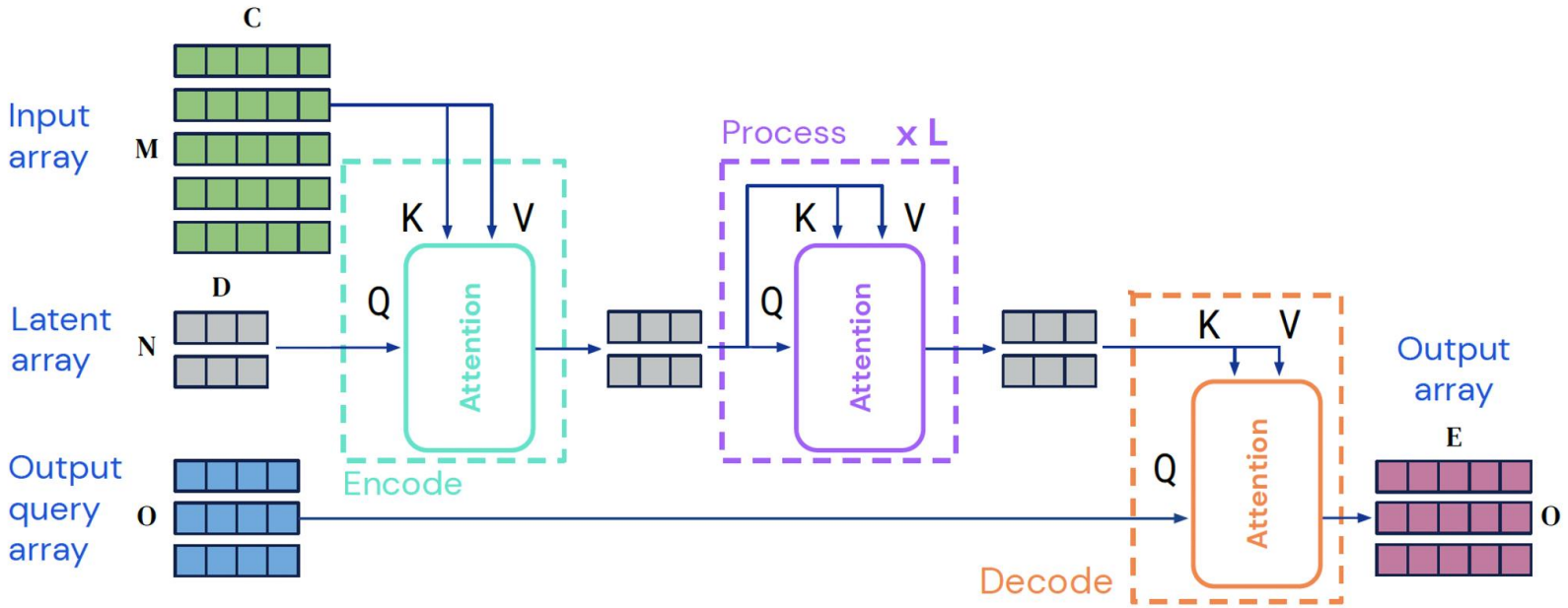
Transformers

[Vaswani et al. Attention is all you need NeurIPS 2017]



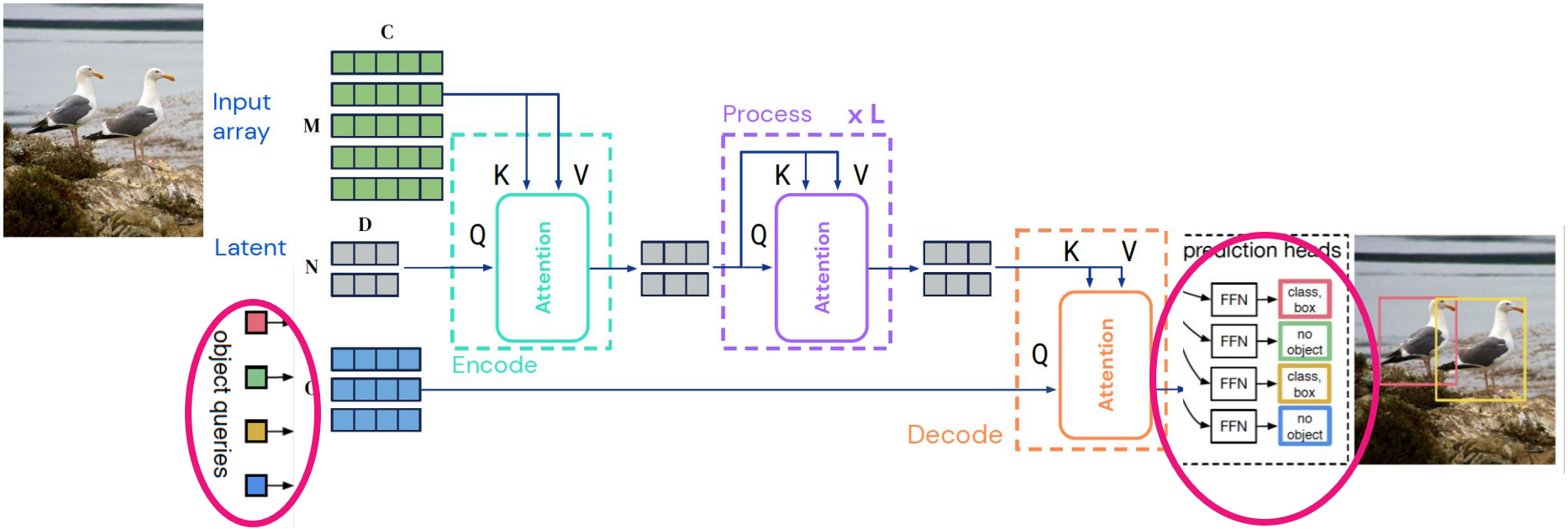
(Visual) Transformers for detection, segmentation, …

Just to complete the big picture
[Perceiver IO A General Architecture for Structured Inputs & Outputs ICLR22]



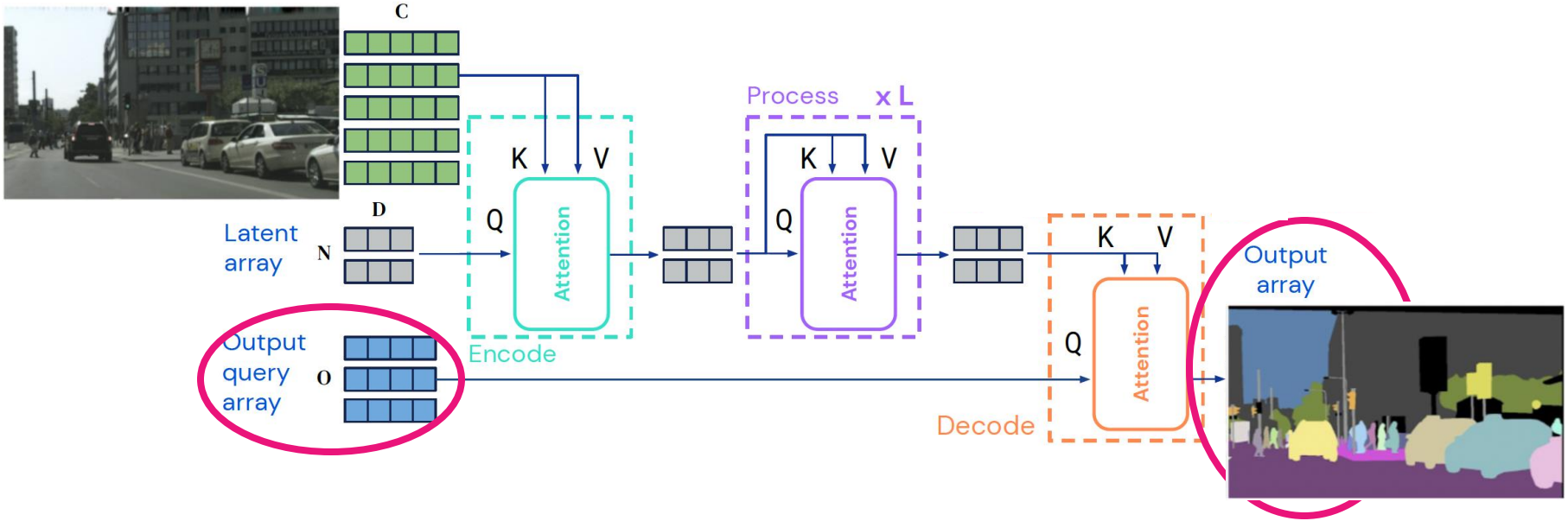
(Visual) Transformers for detection, segmentation, …

Output query array / Output array defines the downstream task: detection



(Visual) Transformers for detection, segmentation, …

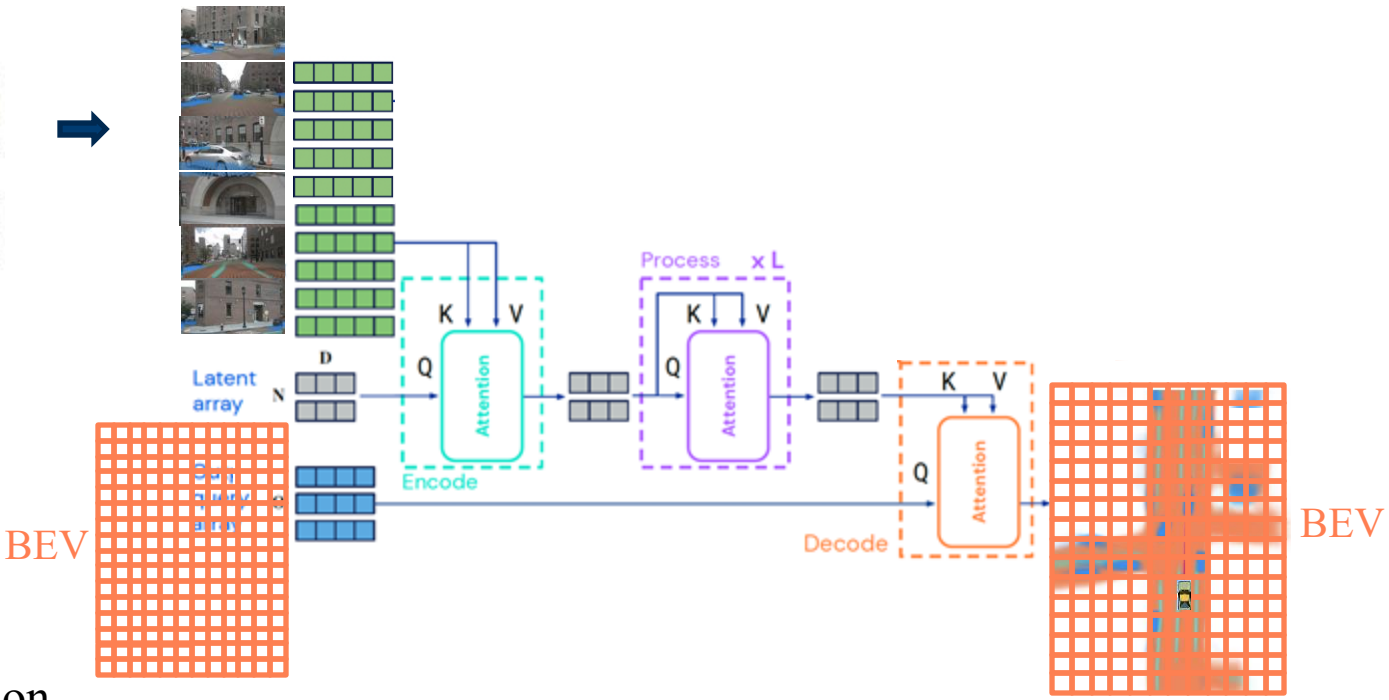
Output query array / Output array defines the downstream task: segmentation …



(Visual) Transformers for detection, segmentation, …

Input array = N cameras

Output array = Bird Eye View (BEV) representation

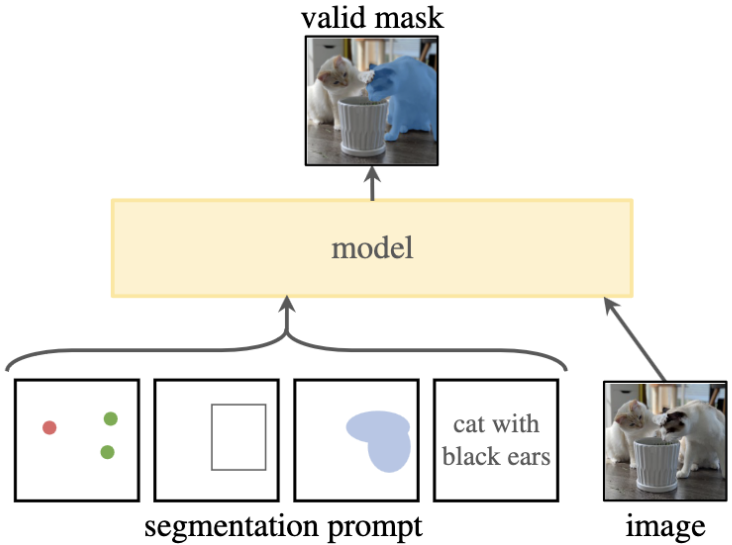


Merging by attention

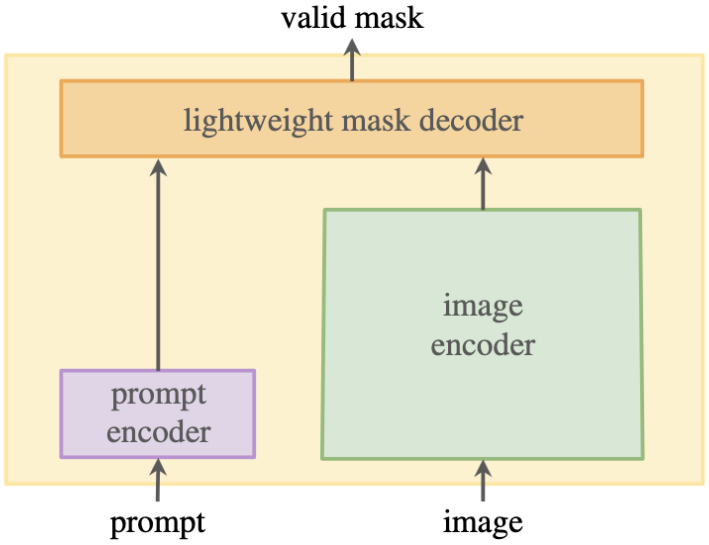
Many Foundation models for Autonomous driving based on this framework

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]



(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)