From Pixels to Preservation: The Power of Large Vision Models in Heritage Content Understanding

Jing Zhang

The University of Sydney

2024.10.28

Content

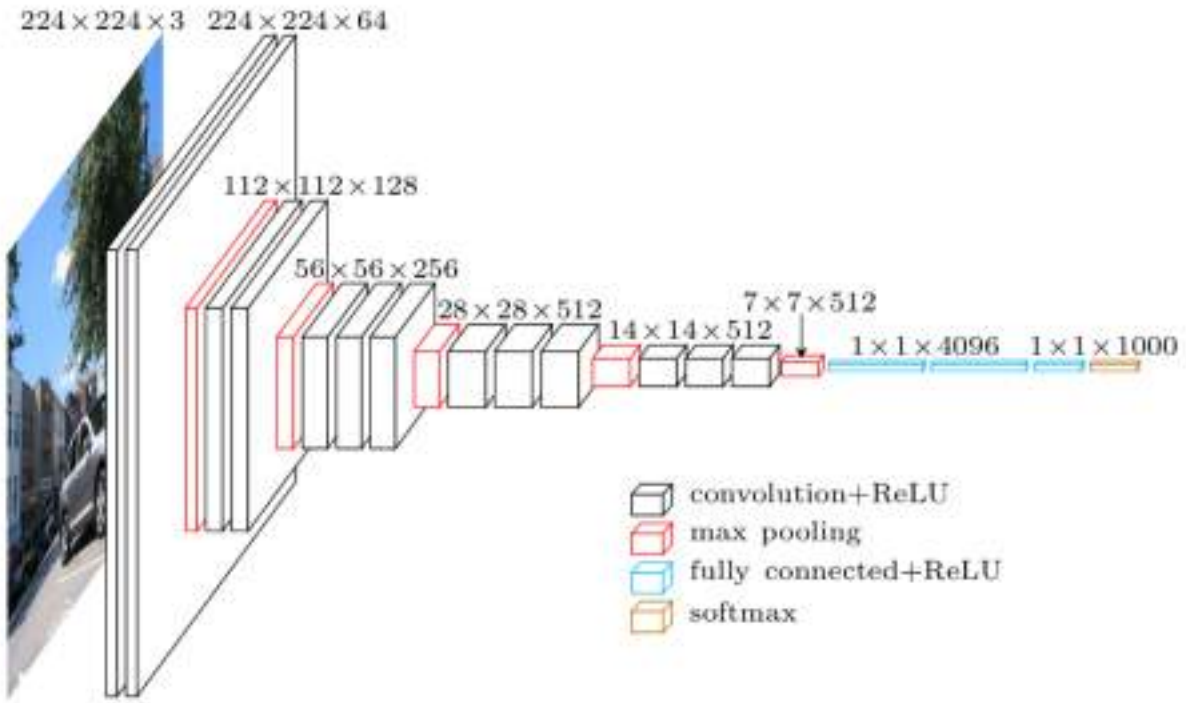
- Foundation Vision Model
  - Architecture Design
  - Model Pretraining
  - Image Generation
- Applications in Heritage Content Understanding
- Outlook

Content

- **Foundation Vision Model**
  - **Architecture Design**
  - Model Pretraining
  - Image Generation
- Applications in Heritage Content Understanding
- Outlook

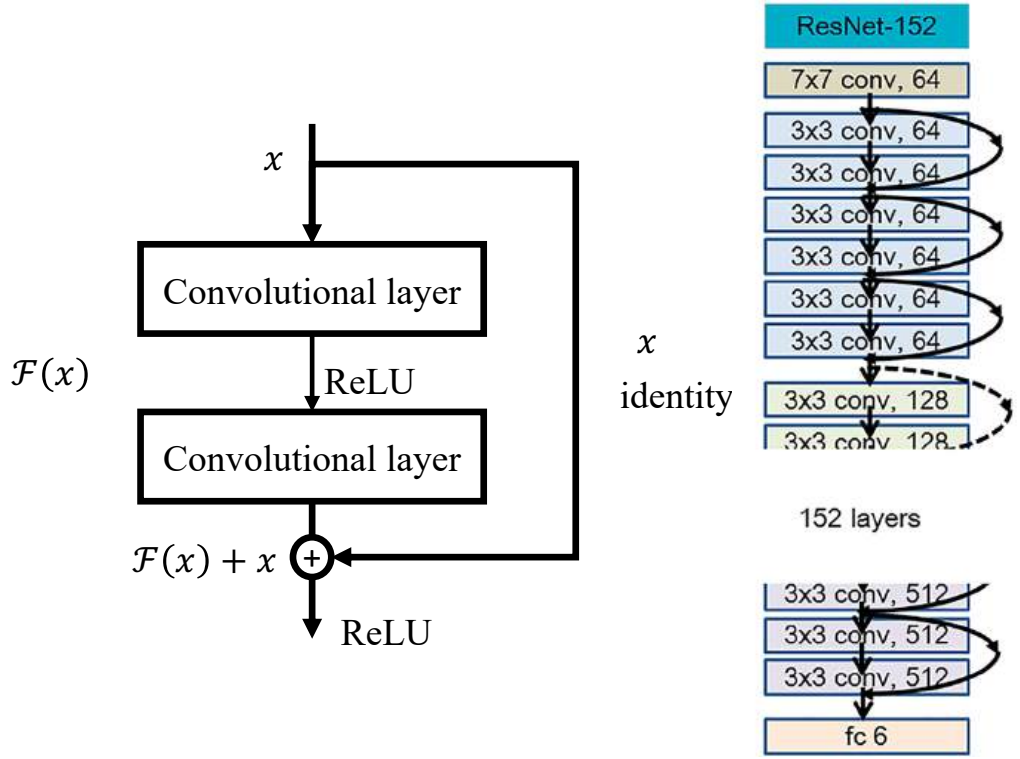
Representative Neural Architectures: CNNs

VGG: one of the most representative plain convnets



- *Stacked 3*3 convolutions*
- *Try to scale the model size*

ResNet: one of the most widely used CNNs

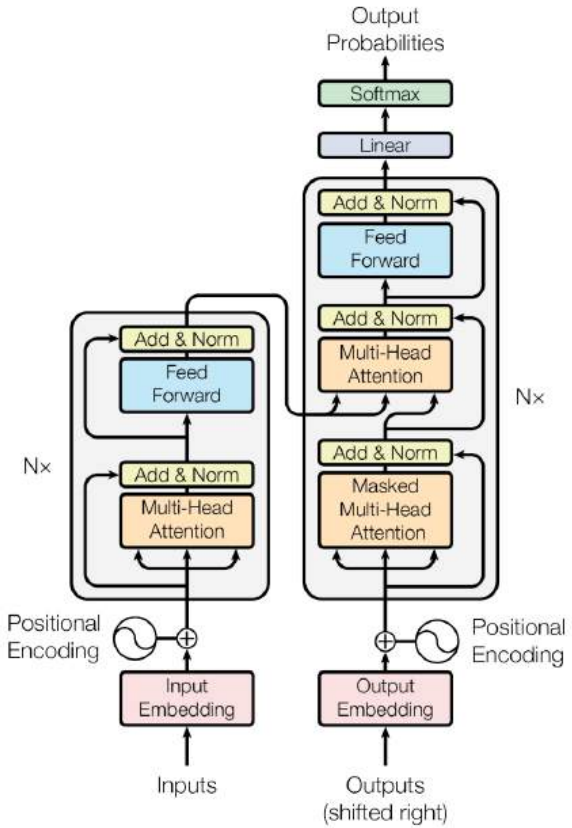


The residual connection ease the training of substantially deeper networks, up to 1000+ layers.

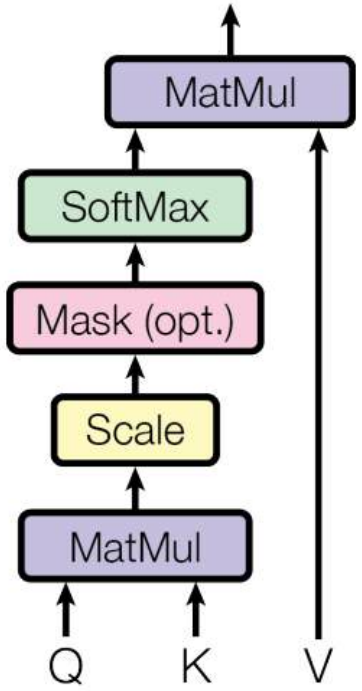
Simonyan & Zisserman: "Very Deep Convolutional Networks for Large-Scale Image Recognition", ICLR 2015.

Representative Neural Architectures: Transformers

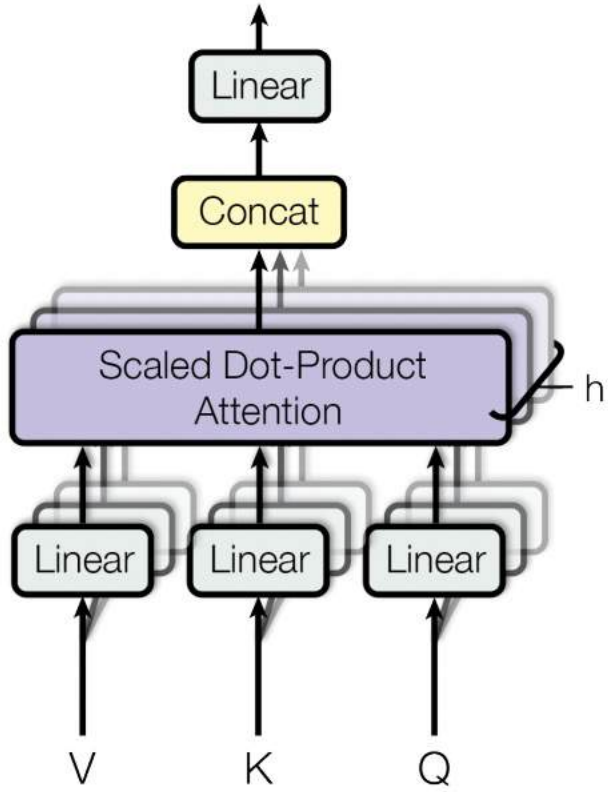
Transformer Architecture



Scaled Dot-Product Attention



Multi-head Attention

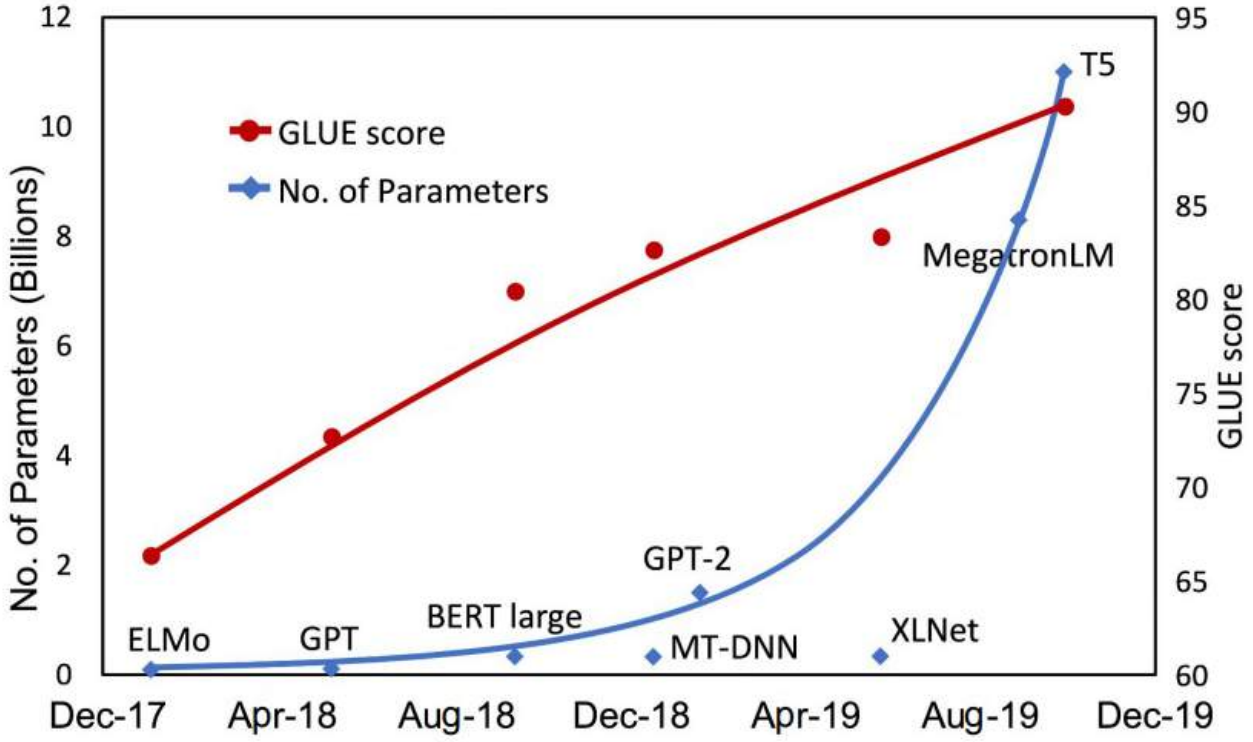


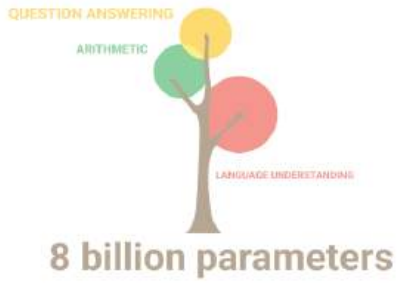
Vaswani, Shazeer, Parmar, et al: "Attention is all you need", NeurIPS 2017.

Representative Neural Architectures: Transformers

The linear performance growth with more model parameters in NLP

- *Showing the potential of transformers with large parameters → Large Language Model*



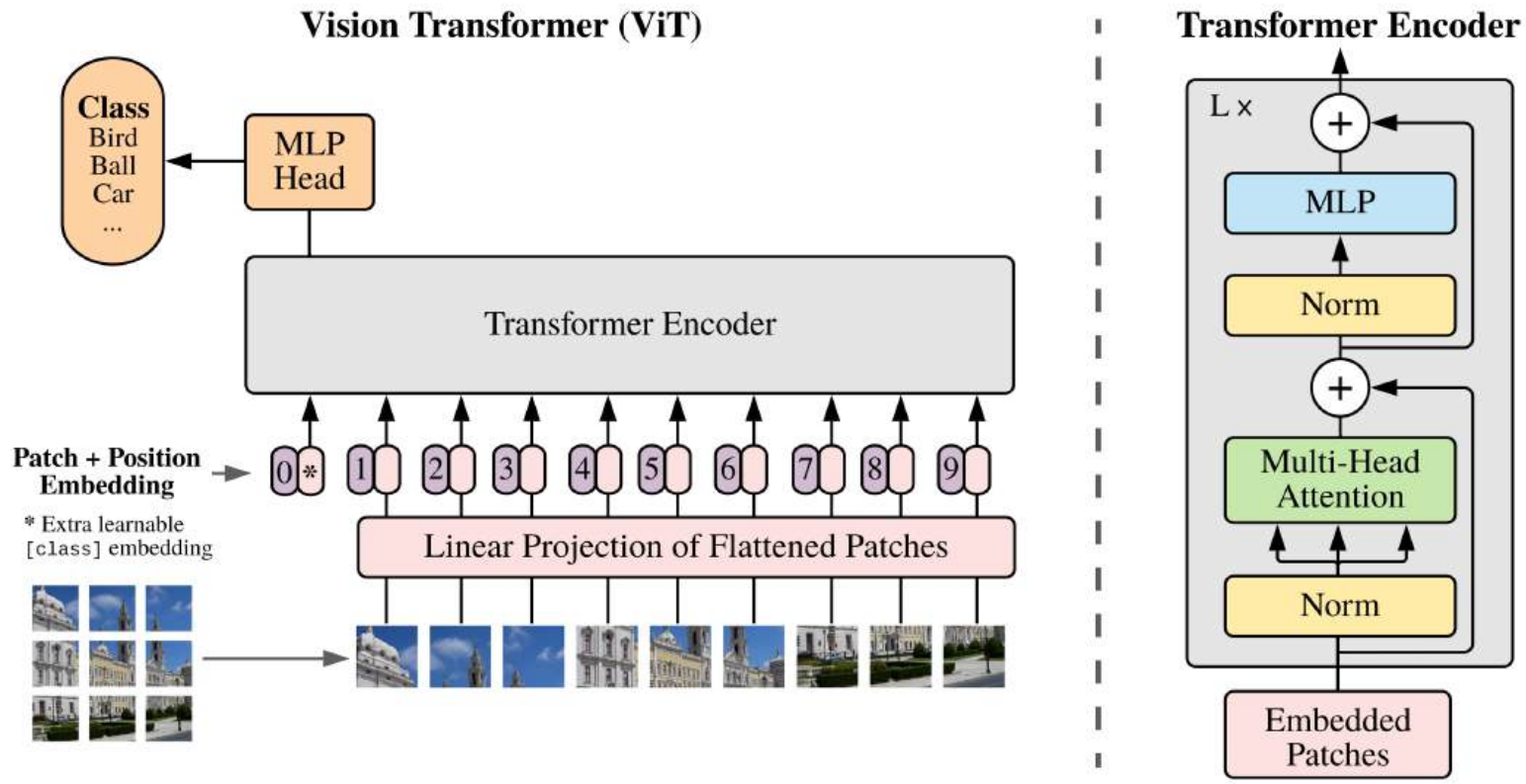


The success of large models in NLP

Vaswani, Shazeer, Parmar, et al: "Attention is all you need", NeurIPS 2017.
Ahmet, Abdulla: "Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach", BigDataSE 2020.

Representative Neural Architectures: Vision Transformers

Transformers for Image Recognition at Scale



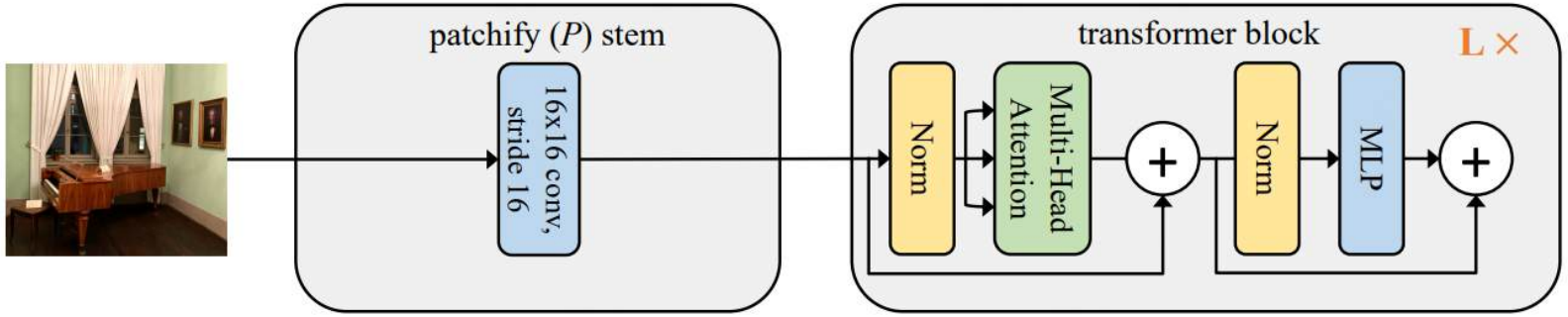
- *Embedding image patches as tokens and treating them as text sequence in NLP*
- *Adding a class token to aggregate information for classification*

Dosovikskiy, Beyer, Kolesnikov, et al.: "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale", ICLR 2020.

Plain Vision Transformers: Improvement over Tokenizer

Early Convolutions

Motivation: large-kernel plus large-stride convolution in pachifying in ViTs runs counter to the typical design in CNNs



Original ViT:
Sensitive to lr and wd choice
Converges slowly
Works with AdamW, but not SGD
Underperforms sota CNNs on ImageNet

ViT with early conv:
Robust to lr and wd choice
Converges quickly
Works with AdamW, and also SGD
Outperforms sota CNNs on ImageNet

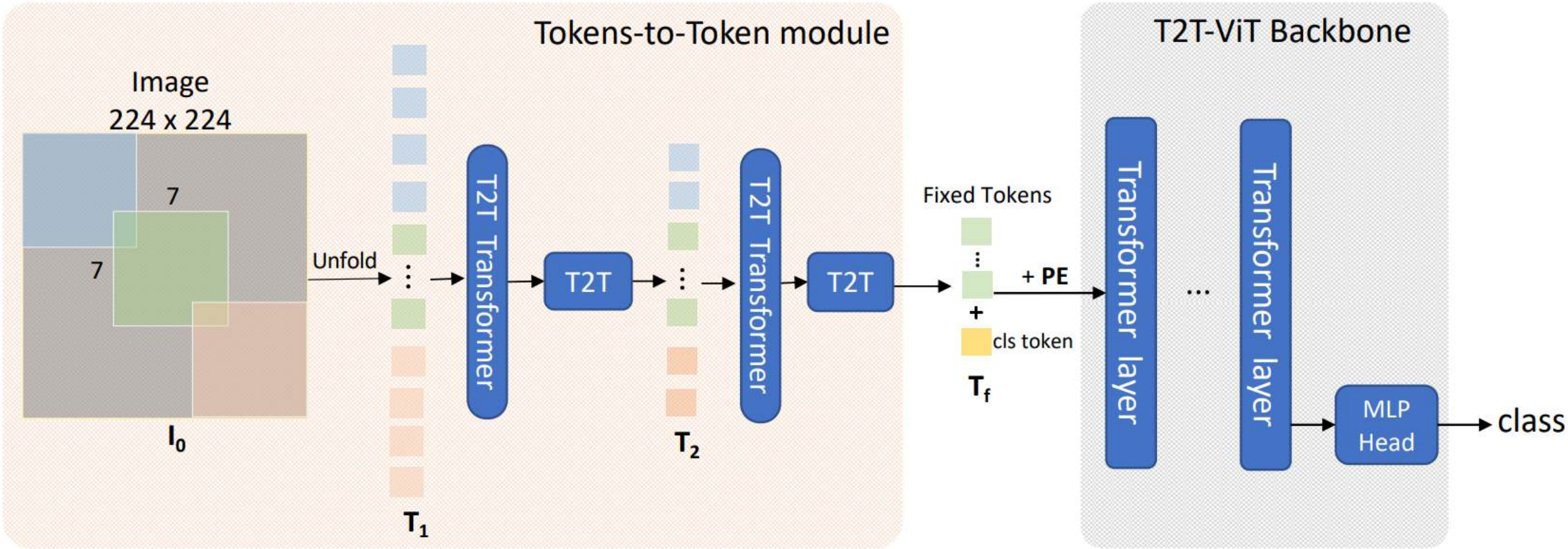
Early convolution captures the low-level edge information better and thus help ViTs see better.

Xiao, Tete, et al. "Early convolutions help transformers see better." NeurIPS 2021.

Plain Vision Transformers: Improvement over Tokenizer

Token to tokens

T2T-ViT model architecture

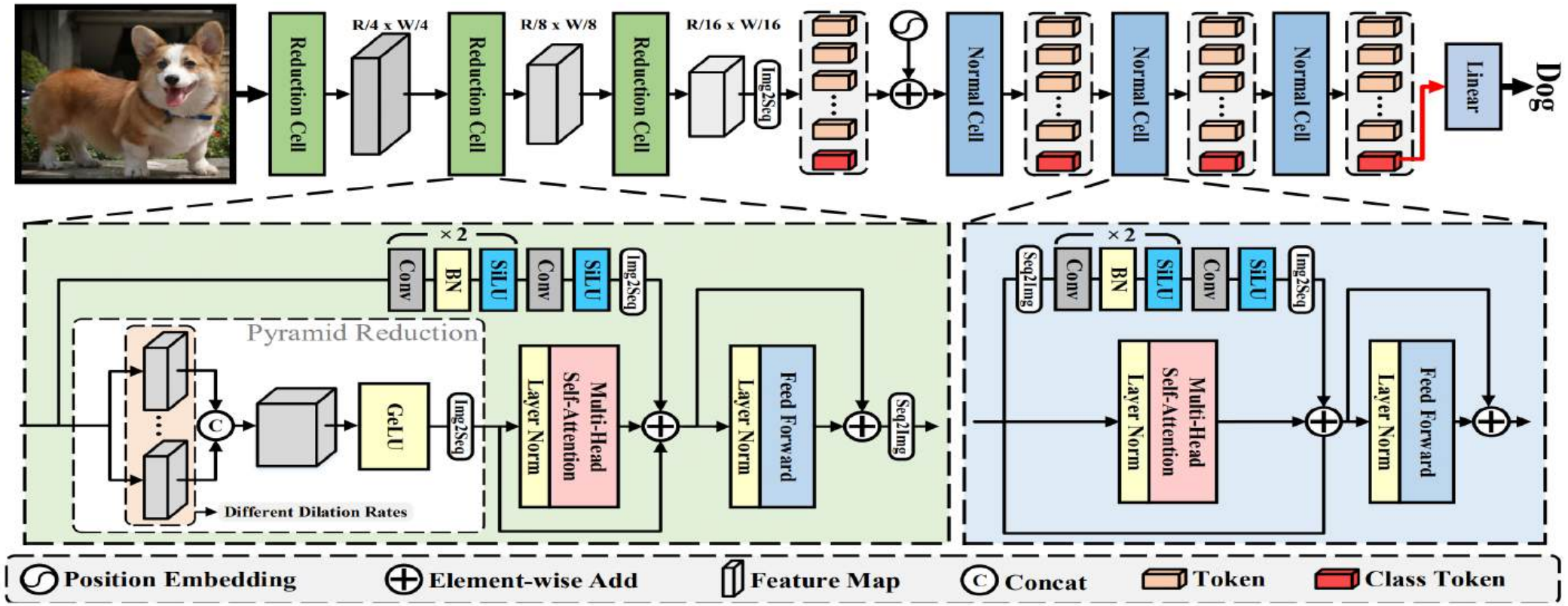


Several tokens-to-tokens modules to **gradually tokenize**.

Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." ICCV 2021.

Plain Vision Transformers: Improvement over Tokenizer

ViTAE Transformer



- Short- and long-range dependencies together.
- Explicitly introduce the intrinsic IBs into transformers with the help of convolutions.
- Gradually tokenizing via stacked RCs for better tokenization.

Xu, Yufei, et al. ”ViTAE: Vision transformer advanced by exploring intrinsic inductive bias." NeurIPS 2021.

Plain Vision Transformers: Improvement over Tokenizer

ViTAE Transformer

ViTAE-H (MAE, 512)	91.2%	644M	Advanced by Exploring Inductive Bias for Image Recognition and Beyond		2022	Transformer
Model soups (ViT-G/14)	91.20%	1843M	Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time		2022	Transformer
Meta Pseudo Labels (EfficientNet-B6-Wide)	91.12%		Meta Pseudo Labels		2020	EfficientNet
MAWS (ViT-6.5B)	91.1%		The effectiveness of MAE pre-pretraining for billion-scale pretraining		2023	
TokenLearner L/8 (24+11)	91.05%	460M	TokenLearner: What Can 8 Learned Tokens Do for Images and Videos?		2021	
Model soups (BASIC-L)	91.03%	2440M	Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time		2022	ALIGN, JFT-3B, Conv+Transformer

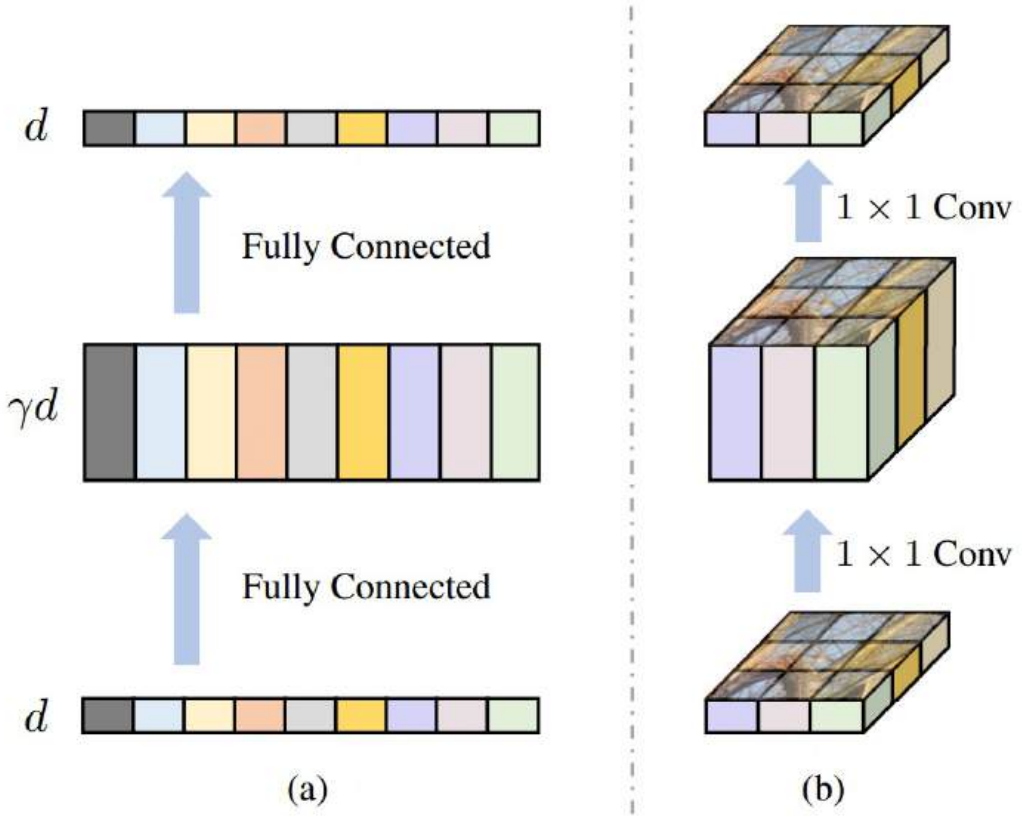
ViTAE-Huge model ranks the first on *ImageNet Real* for image classification

ViTAE Transformer supports various downstream tasks

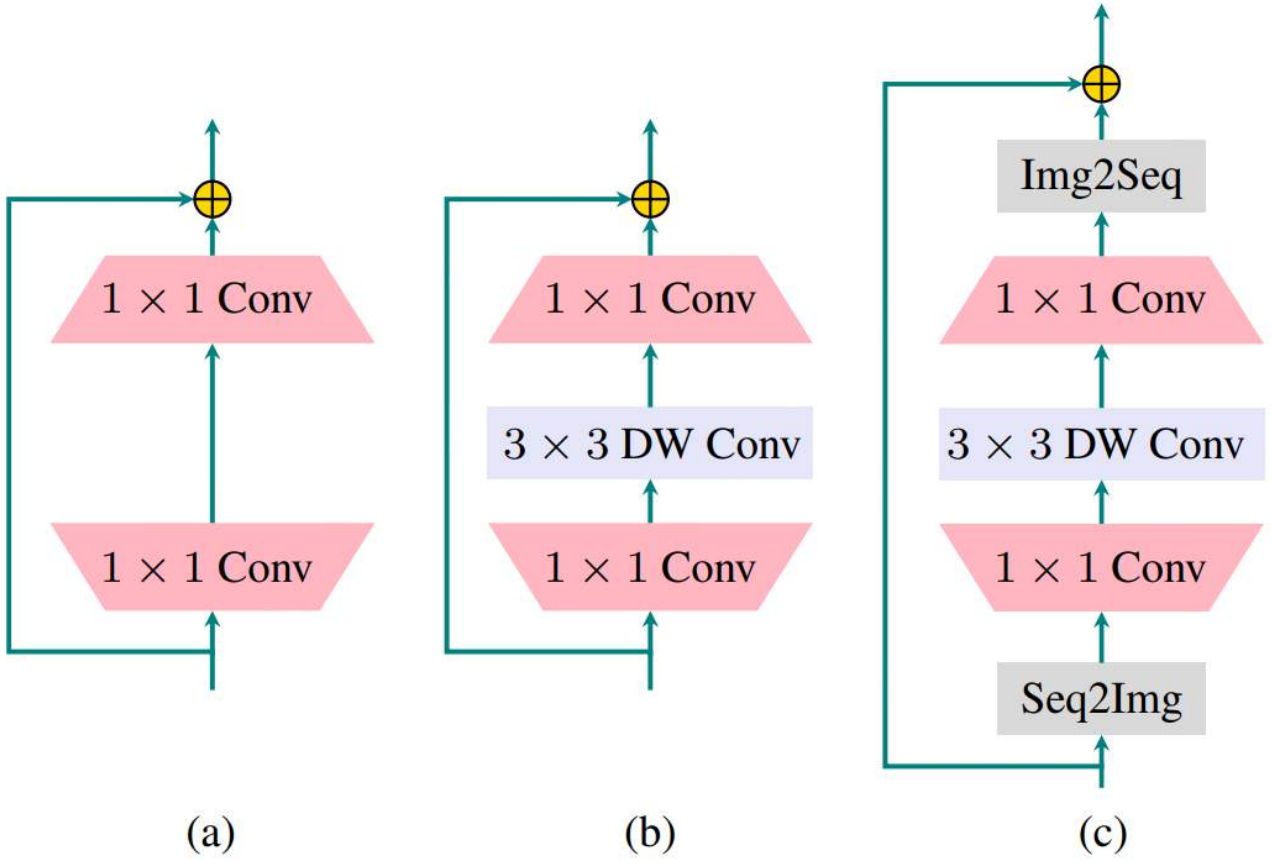
Xu, Yufei, et al. ”ViTAE: Vision transformer advanced by exploring intrinsic inductive bias." NeurIPS 2021.

Plain Vision Transformers: Improvement over MLP

DW-MLP



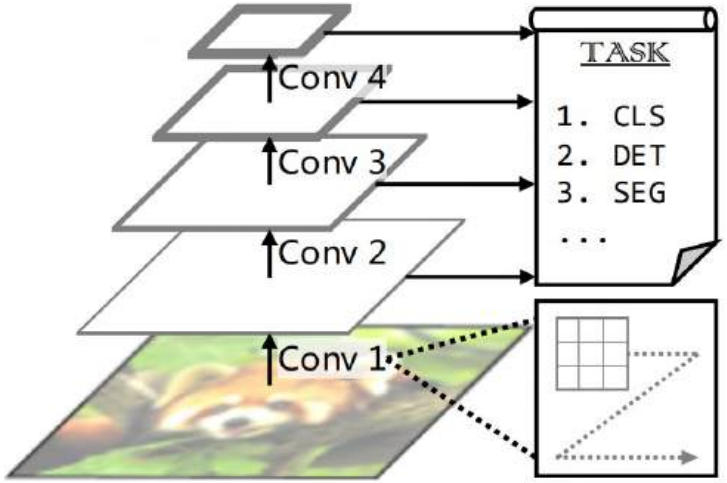
Visualization of the FFN from different perspectives.



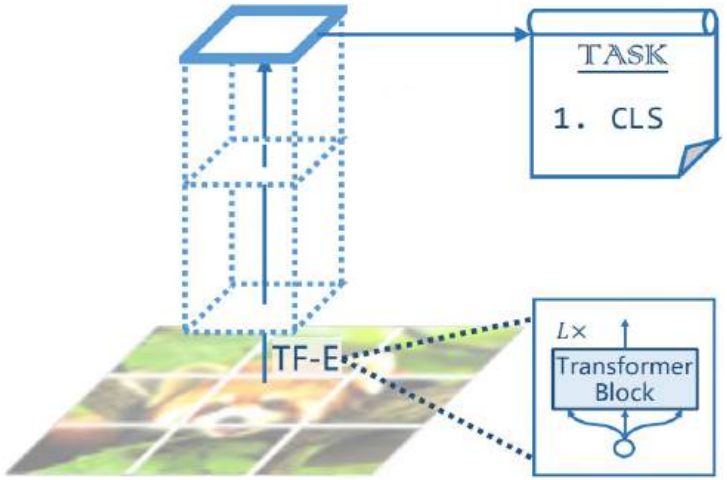
Bringing locality through depth-wise convolution.

Li, Yawei, et al. "Localvit: Bringing locality to vision transformers." CoRR, 2104.05707.

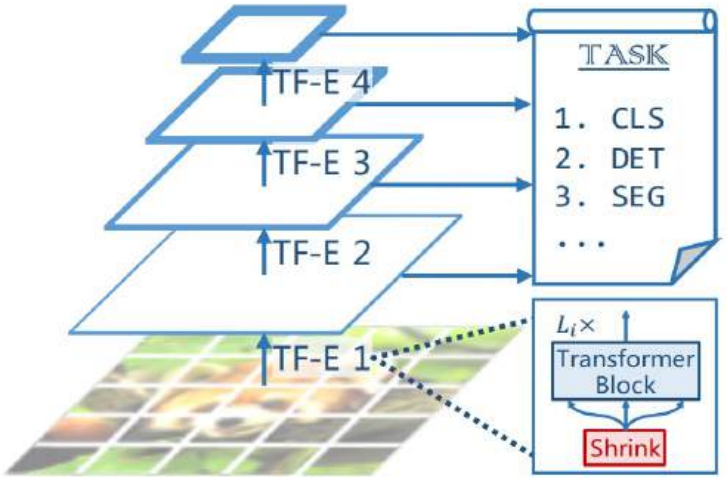
Hierarchical Vision Transformers: Preliminary



(a) CNNs: VGG [54], ResNet [22], *etc.*



(b) Vision Transformer [13]



(c) Pyramid Vision Transformer

Conventional CNNs
- Hierarchical structure with feature pyramid
- Various vision tasks

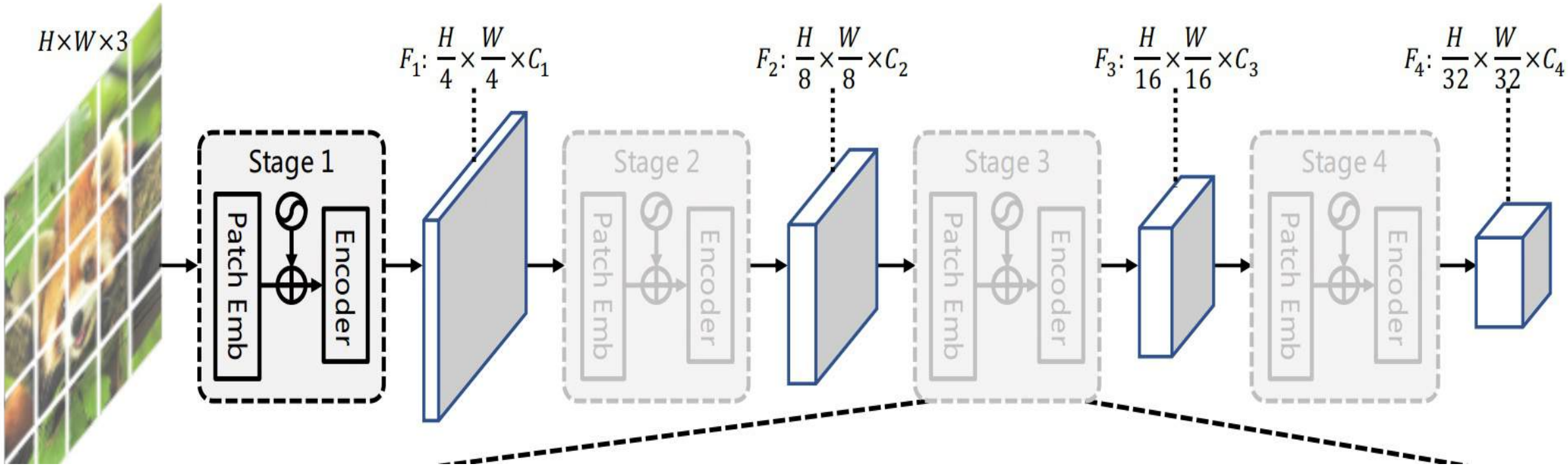
→ Hierarchical ViTs for various vision tasks

ViTs
- Plain structure with all feature maps at the same resolution
- The classification task

Wang, Wenhai, et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." ICCV 2021.

Hierarchical Vision Transformers: Preliminary

Model architecture details

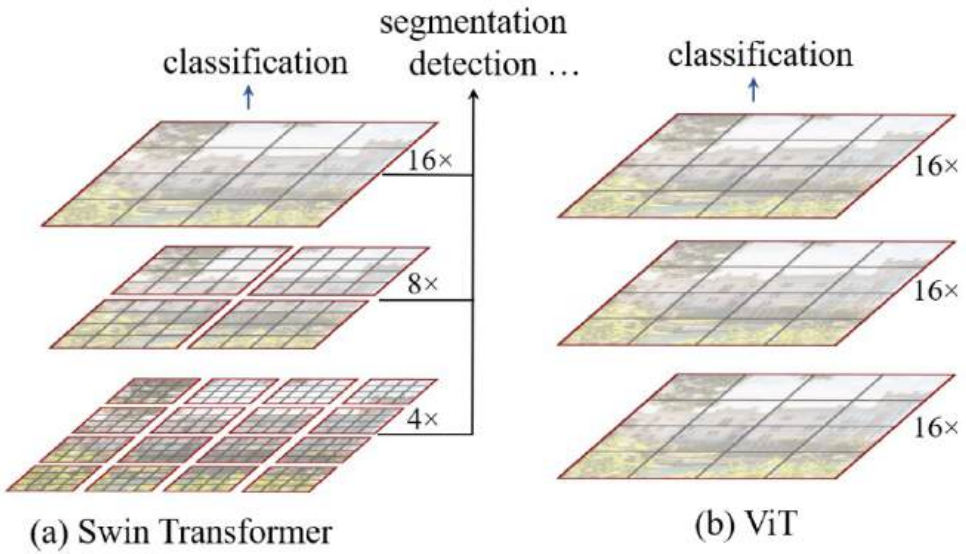


- Four stages in the model with several blocks in each stage.
- The feature map spatial resolution decreases in each stage.
- The output of each stage constructs feature pyramid for various vision tasks.

Wang, Wenhai, et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." ICCV 2021.

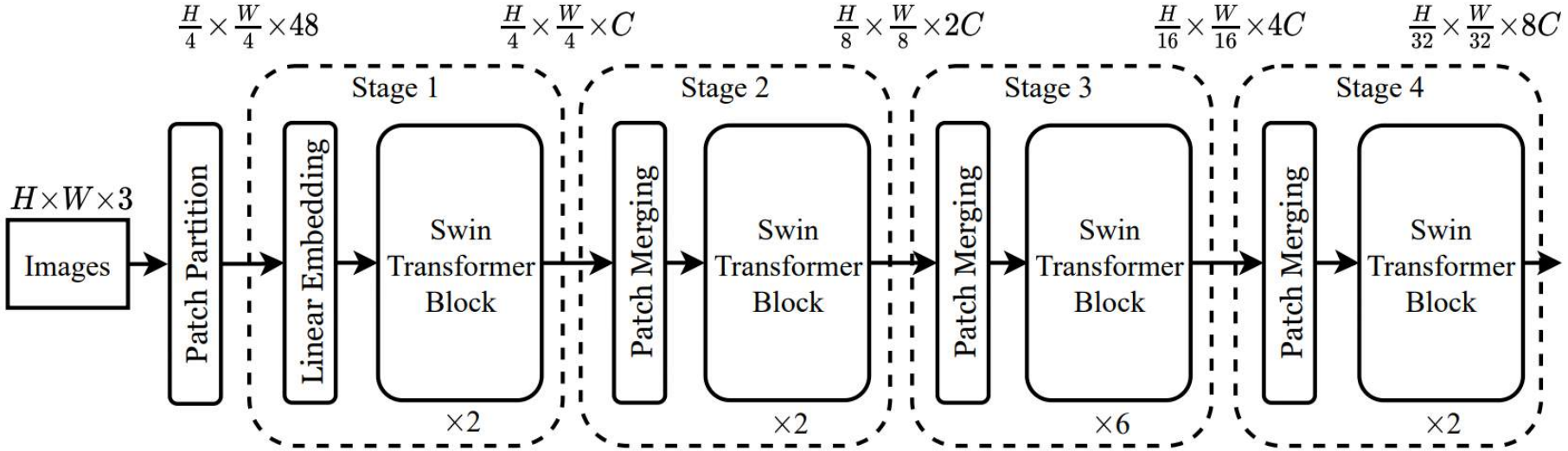
Hierarchical Vision Transformers: Locality

Local window attention



Conventional MHSA calculates over all tokens.
Complexity $O(N^2C)$

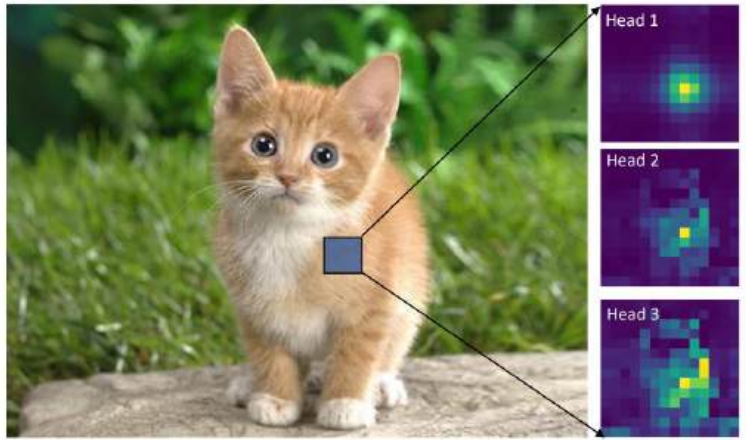
Local window attention calculates within each local window.
Complexity $O(M^2NC)$ where M is the window size and $M \ll N$.

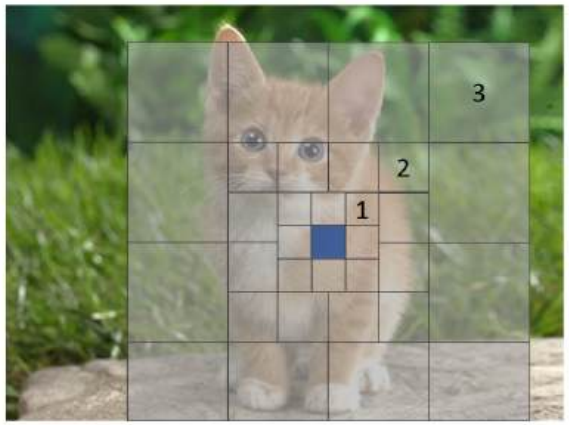


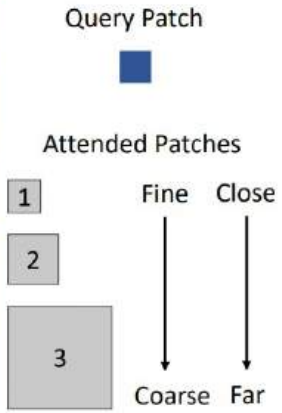
Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." ICCV 2021.

Hierarchical Vision Transformers: Locality

Focal attention

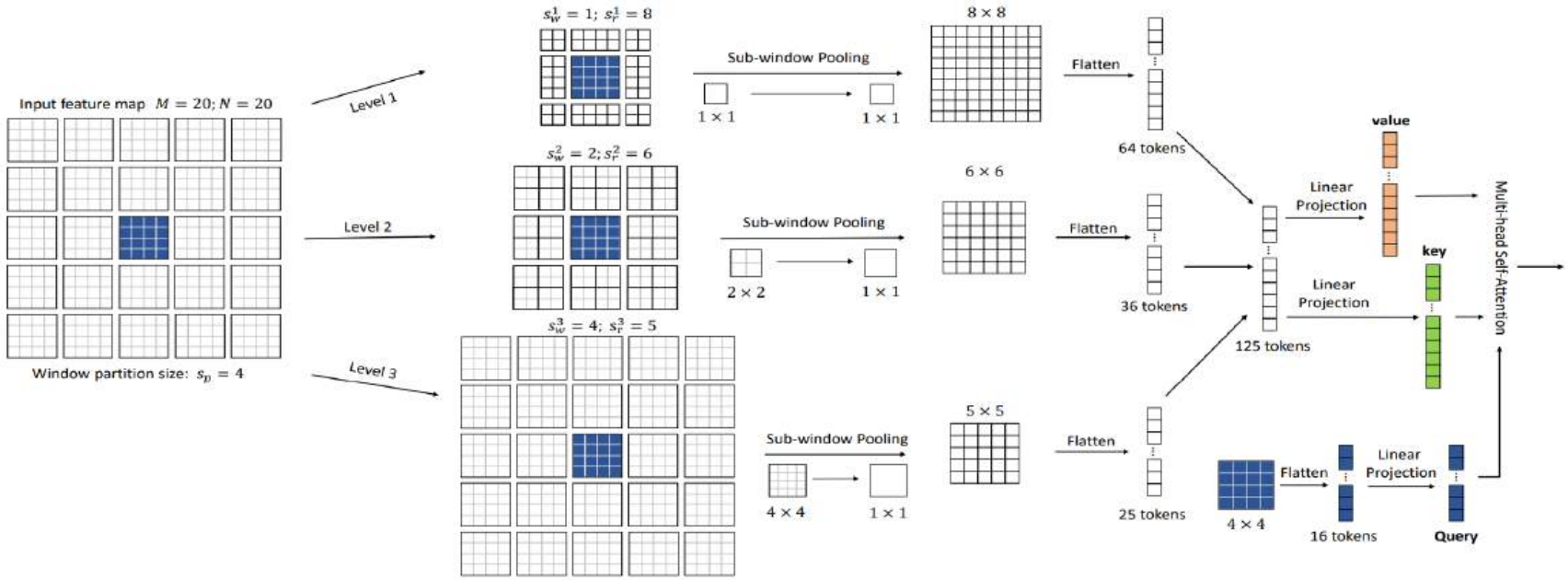






Fine-grained local and coarse-grained global attention.

Enabling better local-global interaction.

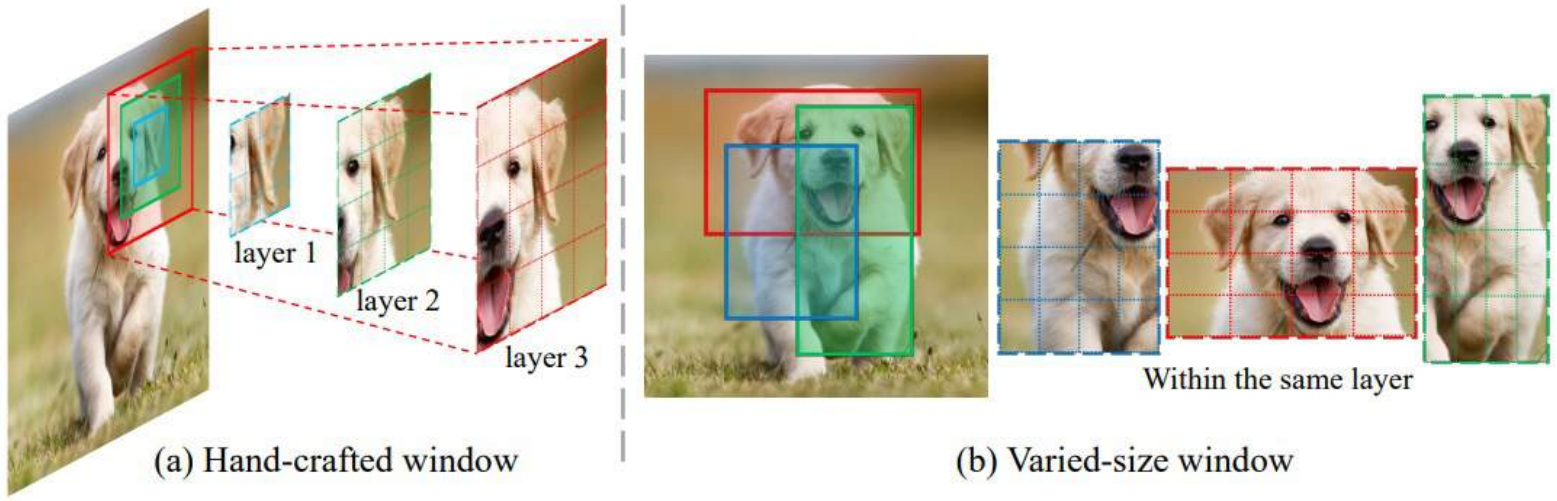


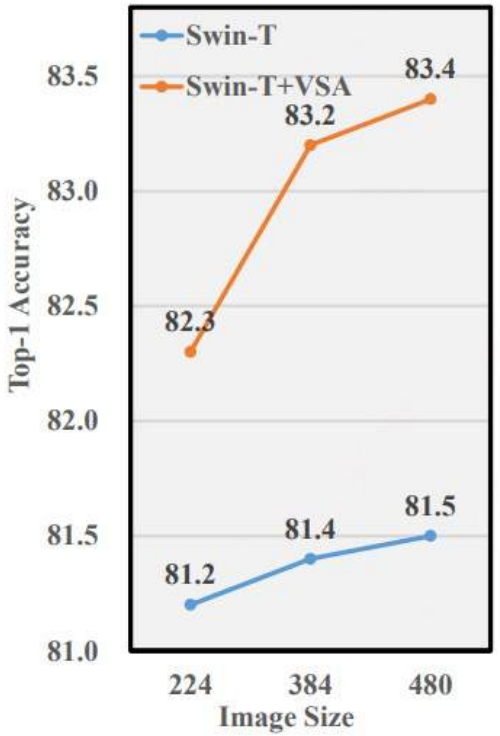
Focal attention examples with three granularity levels.

Yang, Jianwei, et al. "Focal self-attention for local-global interactions in vision transformers." NeurIPS 2021

Hierarchical Vision Transformers: Locality

Varied-size window attention (VSA)





Motivation:
- fixed window size in W-MSA
- modelling long-term dependencies
- modelling objects of different sizes.

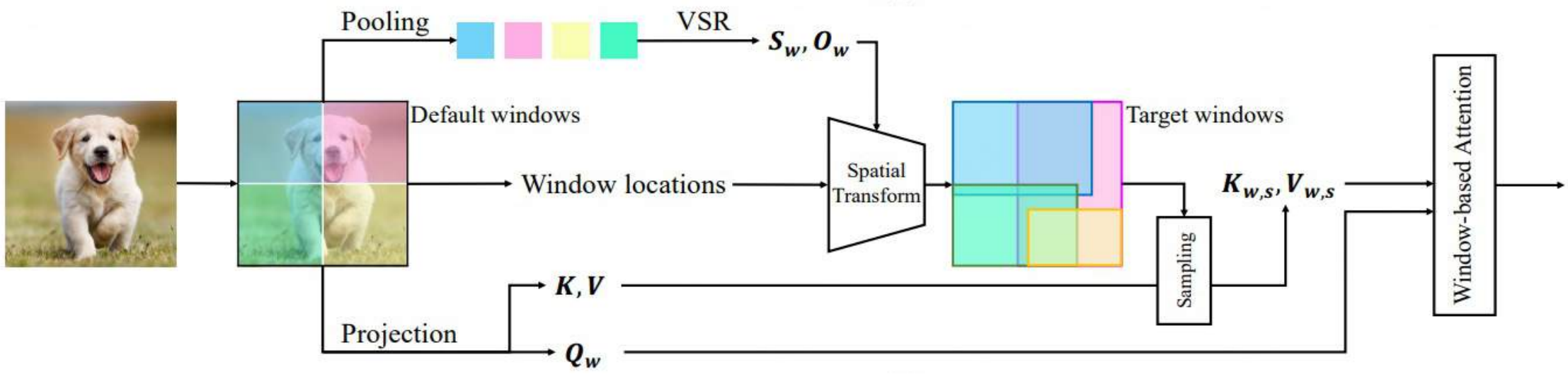
Solution: learnable window configurations enable
- modelling long-range dependencies with large windows
- capturing rich context from diverse windows
- promoting information exchange among windows

Result: VSA
- well adapts to images of various sizes.
- surpasses its baseline by a large margin.

Zhang, Qiming, et al. "VSA: Learning Varied-Size Window Attention in Vision Transformers." ECCV 2022.

Hierarchical Vision Transformers: Locality

Varied-size window attention (VSA)



Motivation:

- fixed window size in W-MSA
- modelling long-term dependencies
- modelling objects of different sizes.

Solution: learnable window configurations enable

- modelling long-range dependencies with large windows
- capturing rich context from diverse windows
- promoting information exchange among windows

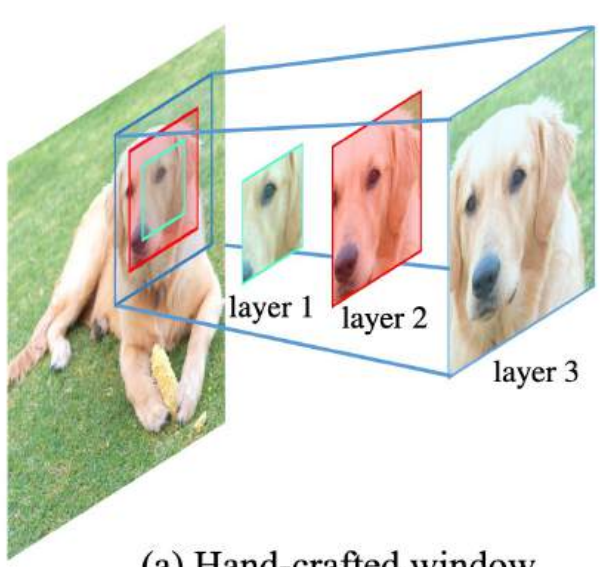
Result: VSA

- well adapts to images of various sizes.
- surpasses its baseline by a large margin.

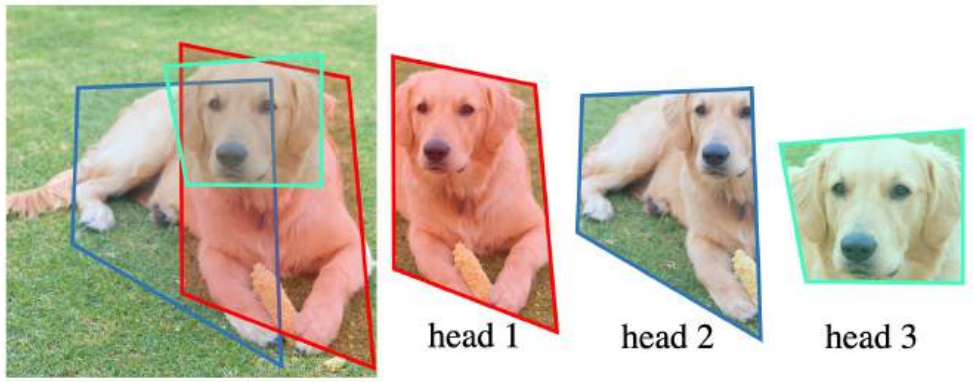
Zhang, Qiming, et al. "VSA: Learning Varied-Size Window Attention in Vision Transformers." ECCV 2022.

Hierarchical Vision Transformers: Locality

Quadrangle attention (QA)



(a) Hand-crafted window



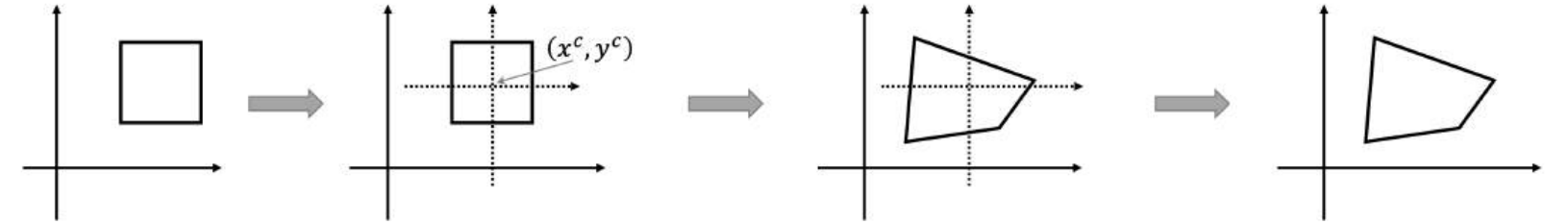
(a) Learned quadrangles

QA learns the sizes, offsets, rotation, shearing, and projection for each default window.

$$T_s = \begin{pmatrix} t_1+1 & 0 & 0 \\ 0 & t_2+1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, T_h = \begin{pmatrix} 1 & t_3 & 0 \\ t_4 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$T_r = \begin{pmatrix} \cos t_5 & -\sin t_5 & 0 \\ \sin t_5 & \cos t_5 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$T_t = \begin{pmatrix} 1 & 0 & \beta_1 t_6 \\ 0 & 1 & \beta_2 t_7 \\ 0 & 0 & 1 \end{pmatrix}, T_p = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t_8 & t_9 & 1 \end{pmatrix},$$



(a) A default window (b) Getting relative coordinates (c) Projective transformation upon relative coordinates (d) Coordinates recovery

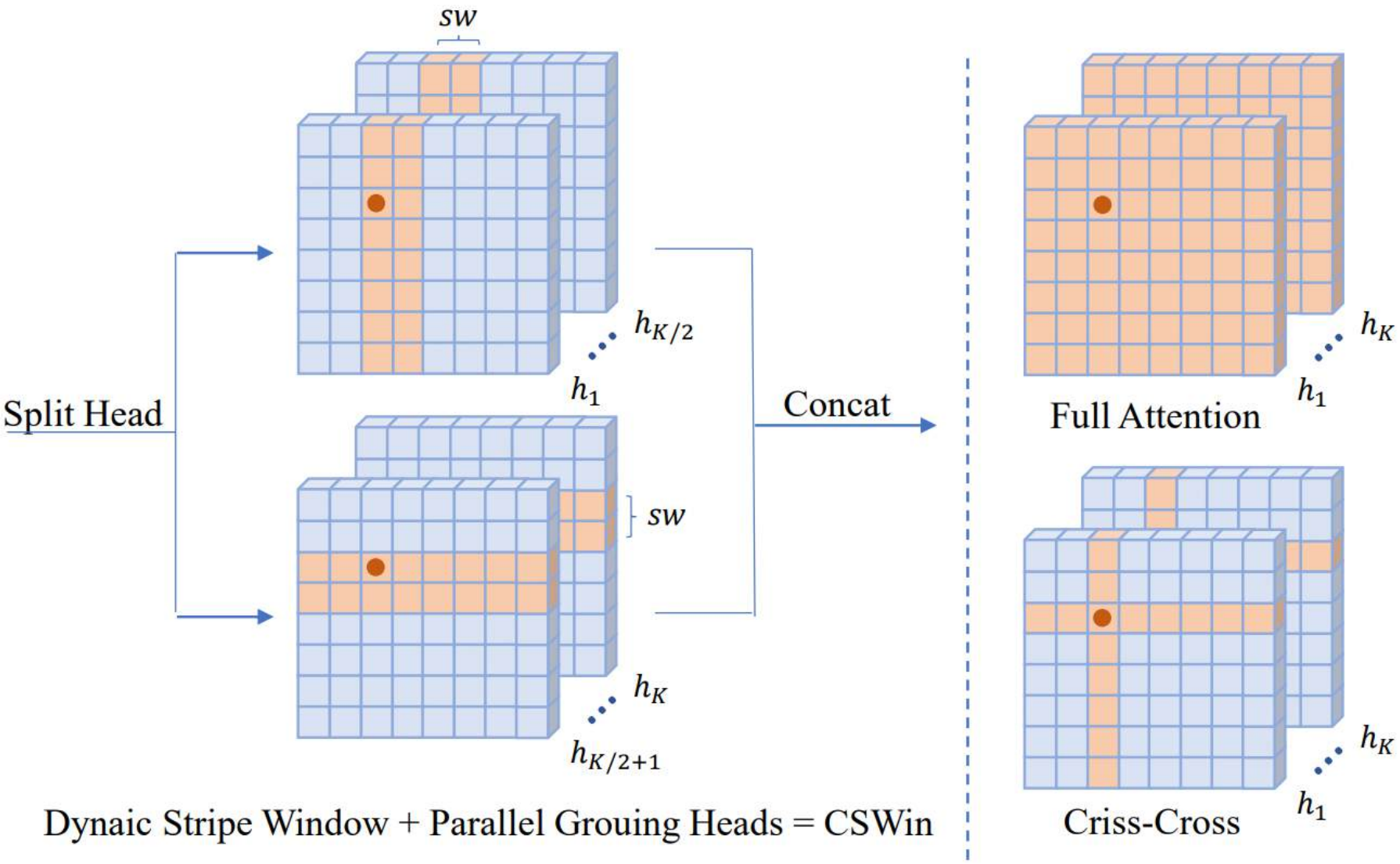
QA transforms windows to quadrangles using projective transformation.

$$T = T_s \times T_h \times T_r \times T_t \times T_p.$$

Qiming Zhang, et al.: Vision Transformer with Quadrangle Attention. IEEE TPAMI 2023.

Hierarchical Vision Transformers: Locality

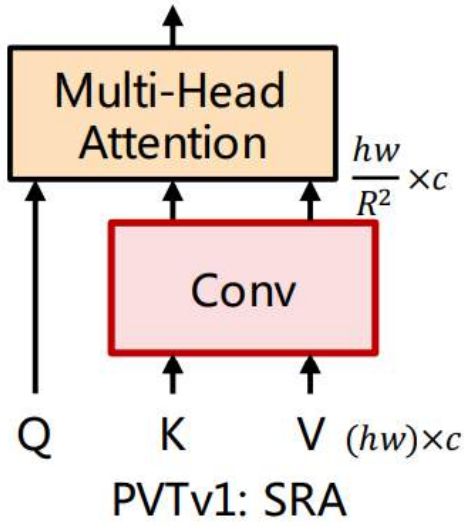
Cross-Shaped Windows Attention



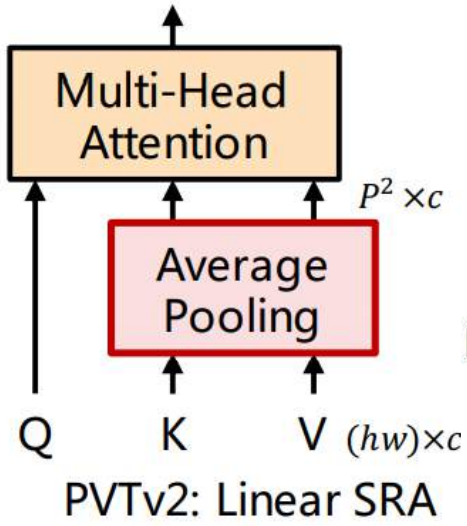
Xiaoyi Dong, et al.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. CVPR 2022.

Hierarchical Vision Transformers: Spatial Redundancy in Self-attention

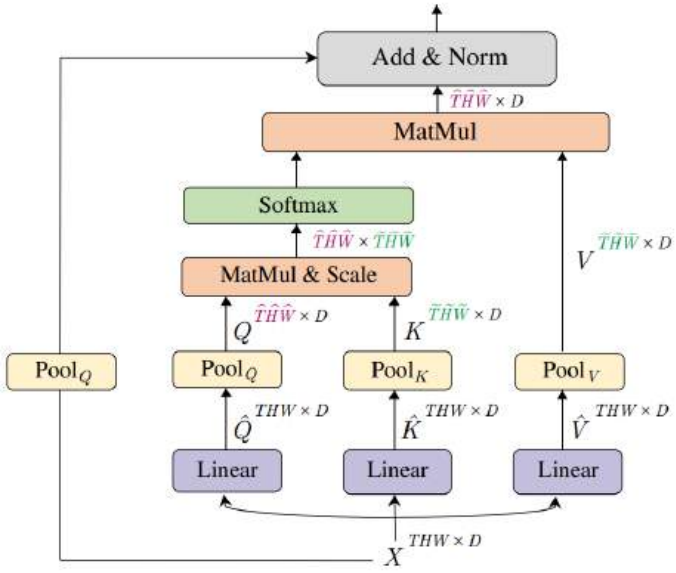
Representative methods



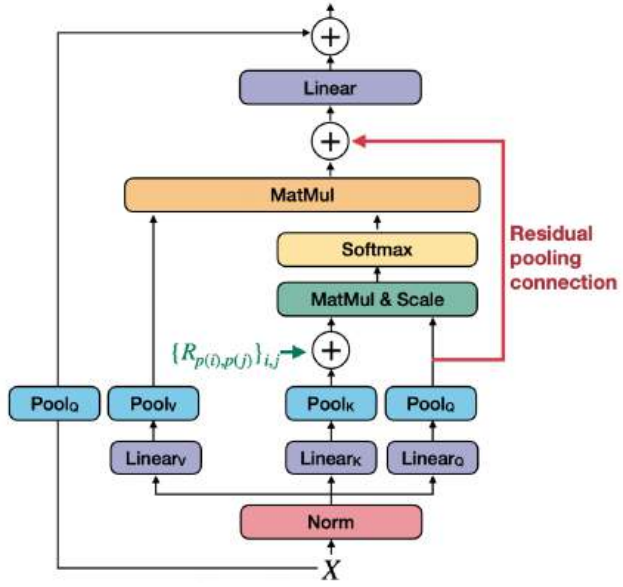
PVTv1: SRA



PVTv2: Linear SRA



MViT



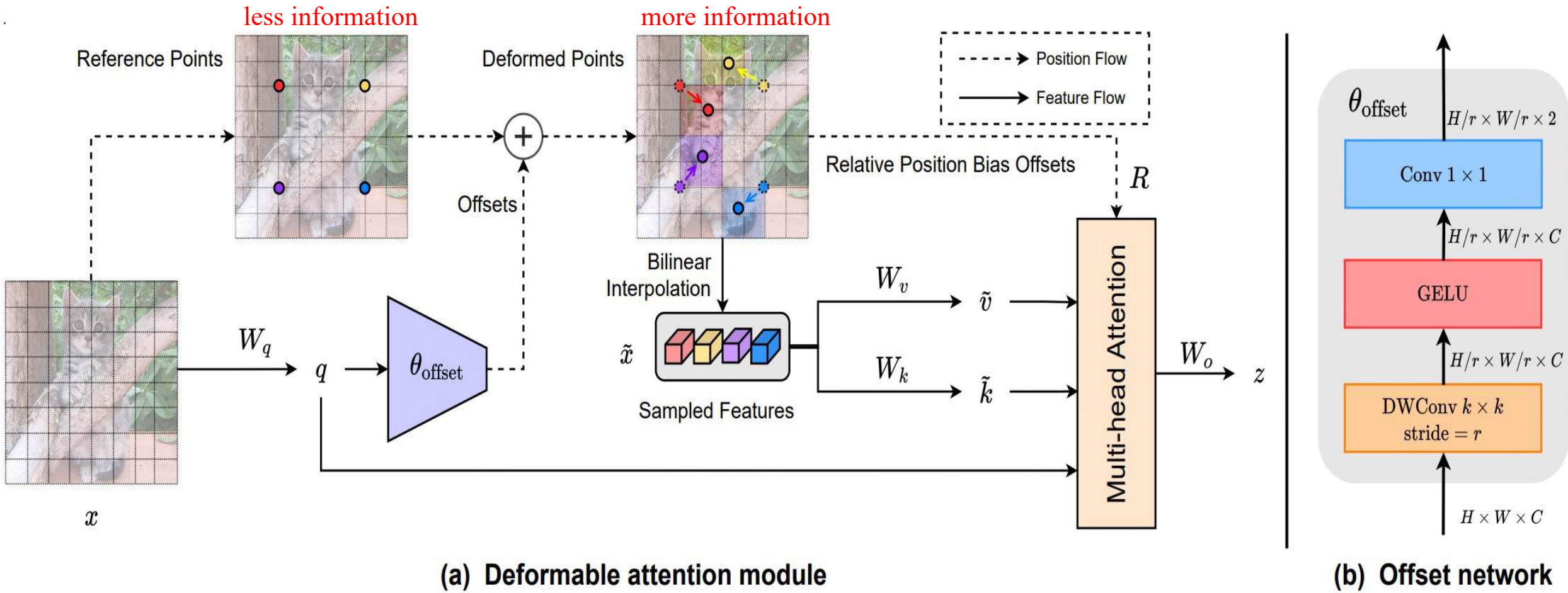
MViTv2

- Pooling attention in MViT enables to downsample feature maps.

Wang, Wenhai, et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." ICCV 2021.
Wang, Wenhai, et al. "Pvt v2: Improved baselines with pyramid vision transformer." Computational Visual Media 2022.
Fan, Haoqi, et al. "Multiscale vision transformers." ICCV 2021.
Li, Yanghao, et al. "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection." CVPR 2022.

Hierarchical Vision Transformers: Spatial Redundancy in Self-attention

Deformable attention



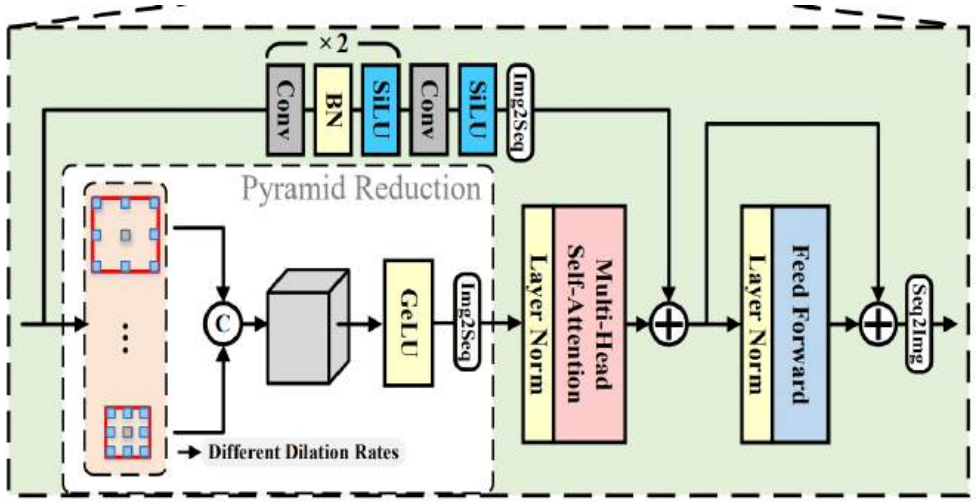
(a) Deformable attention module **(b) Offset network**

- Shifted window attention in the bottom two stages and DAT in the top two stages.
- Controlling the reference point number to balance the computation complexity.

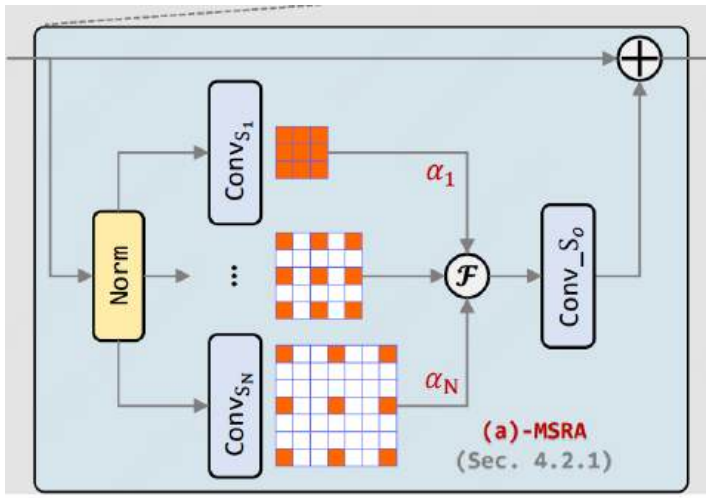
Xia, Zhuofan, et al. "Vision transformer with deformable attention." CVPR 2022.

Hierarchical Vision Transformers: Multi-scale ViTs

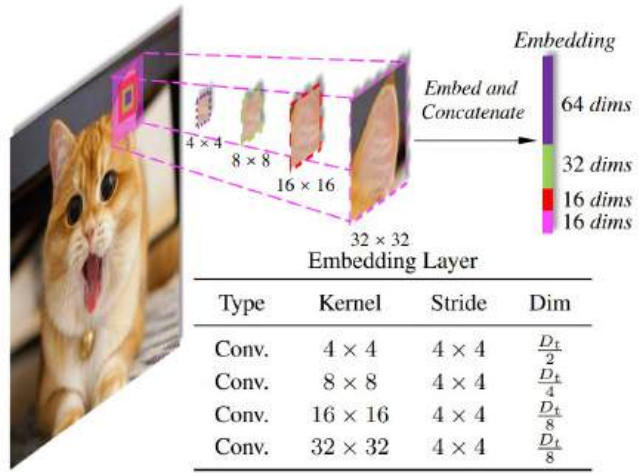
Convolution with **different kernel sizes** to capture multi-scale information



ViTAEv2 [Zhang et al. 2022]

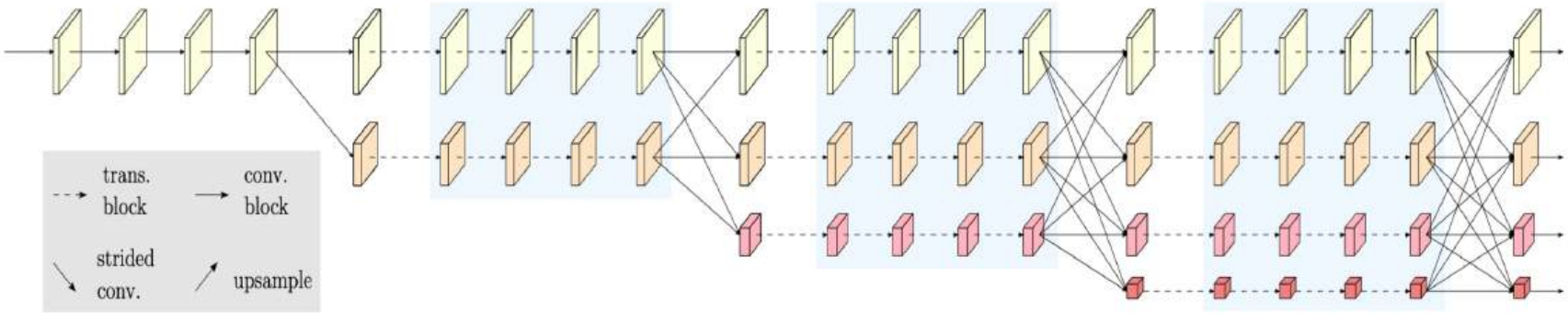


EATFormer [Zhang et al. 2022]



Type	Kernel	Stride	Dim
Conv.	4 × 4	4 × 4	$\frac{D_1}{2}$
Conv.	8 × 8	4 × 4	$\frac{D_1}{4}$
Conv.	16 × 16	4 × 4	$\frac{D_1}{8}$
Conv.	32 × 32	4 × 4	$\frac{D_1}{8}$

CrossFormer [Wang et al. 2021]



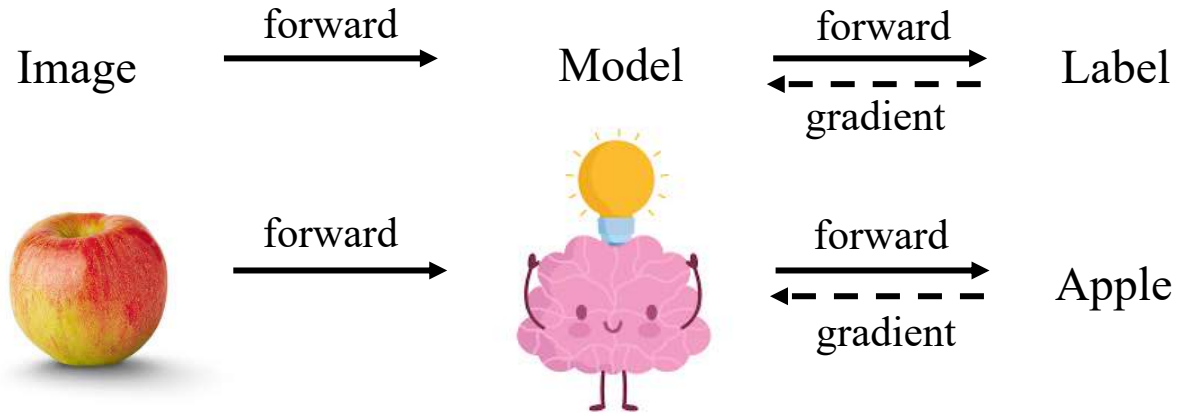
HRFormer [Yuan et al.]

Zhang, Qiming, et al. "Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond." IJCV 2022.
Zhang, Jiangning, et al. "EatFormer: improving vision transformer inspired by evolutionary algorithm." arXiv preprint arXiv:2206.09325, 2022.
Wang, Wenxiao, et al. "Crossformer: A versatile vision transformer hinging on cross-scale attention." ICLR 2021.
Yuan, Y., et al. "HRFormer: High-Resolution Transformer for Dense Prediction." NeurIPS 2021.

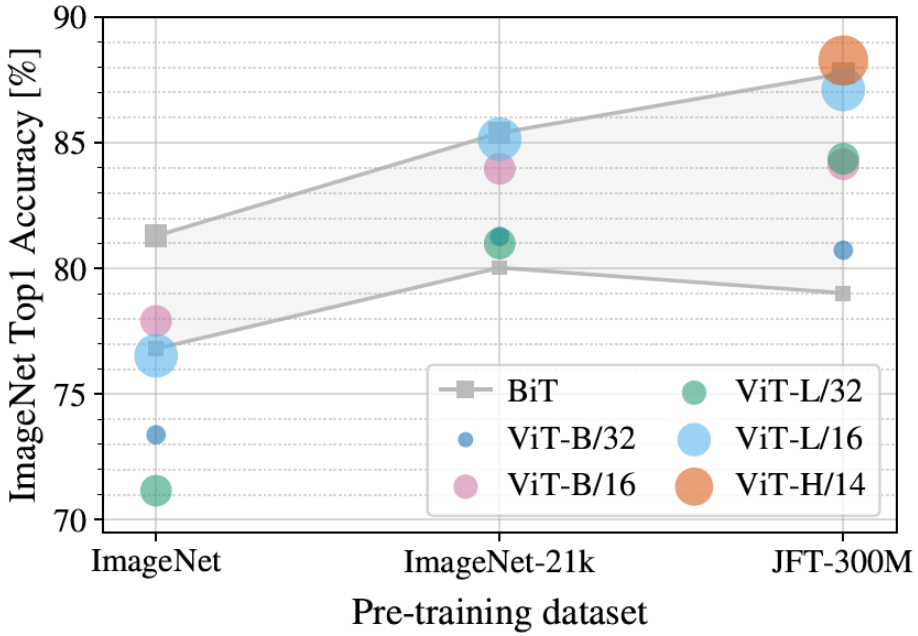
Content

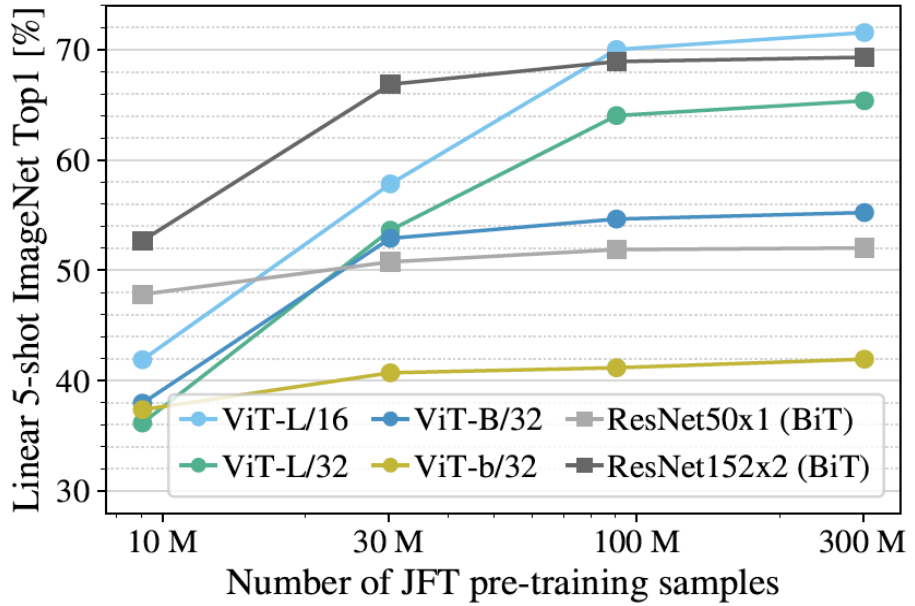
- **Foundation Vision Model**
  - Architecture Design
  - **Model Pretraining**
  - Image Generation
- Applications in Heritage Content Understanding
- Outlook

Supervised Learning



Supervised learning process





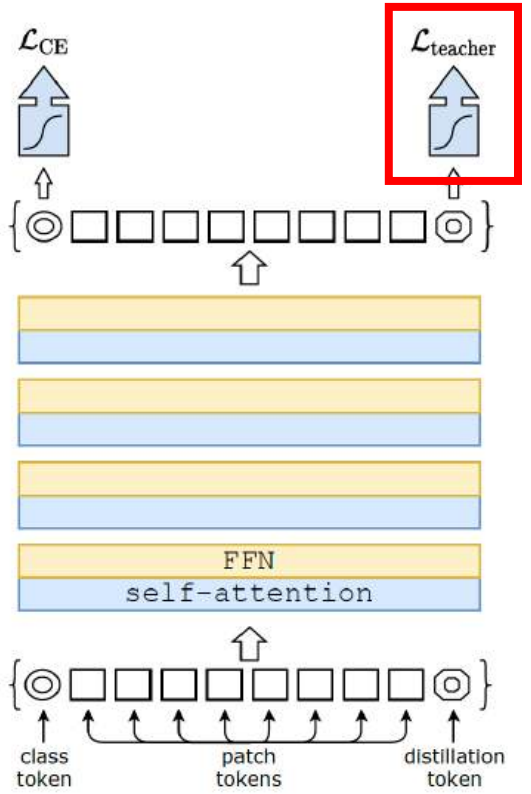
Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." , ICLR 2021.

Supervised Learning

Solution: Data Augmentation & Knowledge Distillation

Methods	ViT-B [15]	DeiT-B
Epochs	300	300
Batch size	4096	1024
Optimizer	AdamW	AdamW
learning rate	0.003	$0.0005 \times \frac{\text{batchsize}}{512}$
Learning rate decay	cosine	cosine
Weight decay	0.3	0.05
Warmup epochs	3.4	5
Label smoothing ε	✗	0.1
Dropout	0.1	✗
Stoch. Depth	✗	0.1
Repeated Aug	✗	✓
Gradient Clip.	✓	✗
Rand Augment	✗	9/0.5
Mixup prob.	✗	0.8
Cutmix prob.	✗	1.0
Erasing prob.	✗	0.25

Augmentation strategy in supervised learning

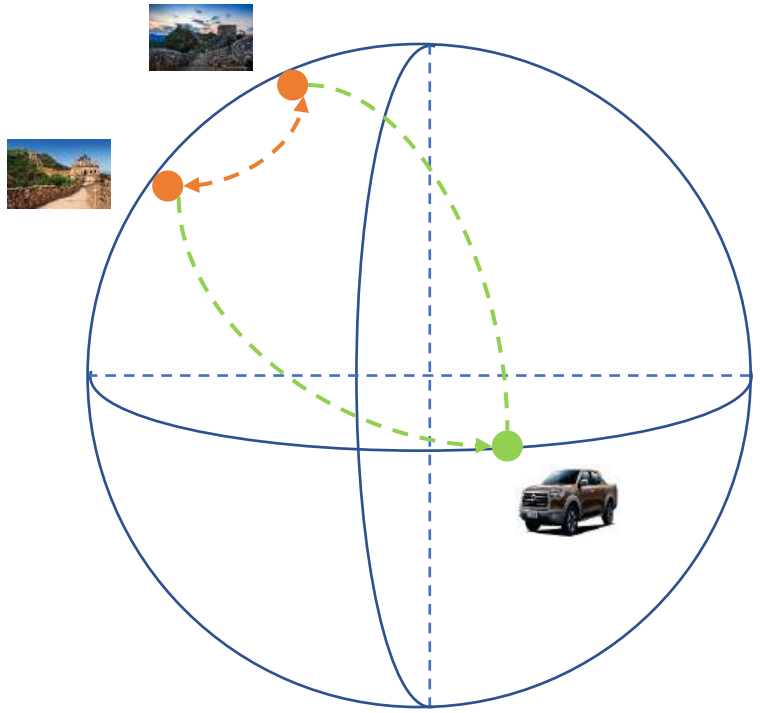


Structure with distillation

Touvron, Hugo, et al. "Training data-efficient image transformers & distillation through attention.", ICML 2021.

Self-Supervised Learning

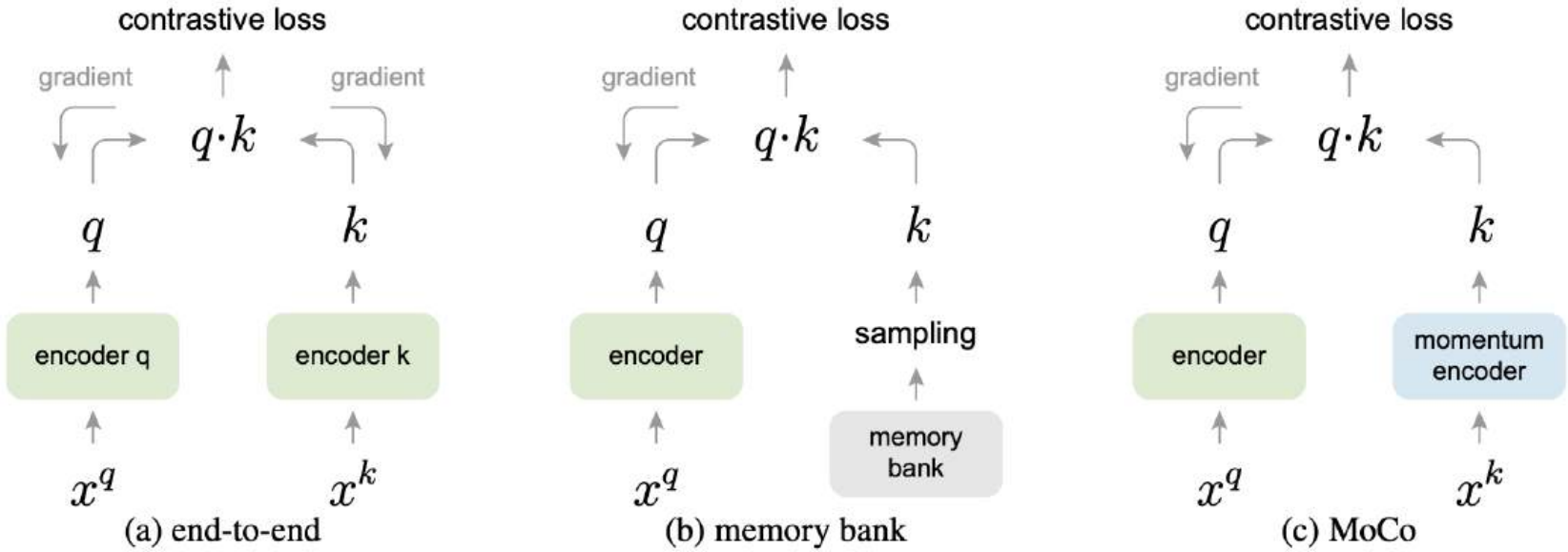
Review Contrastive Learning



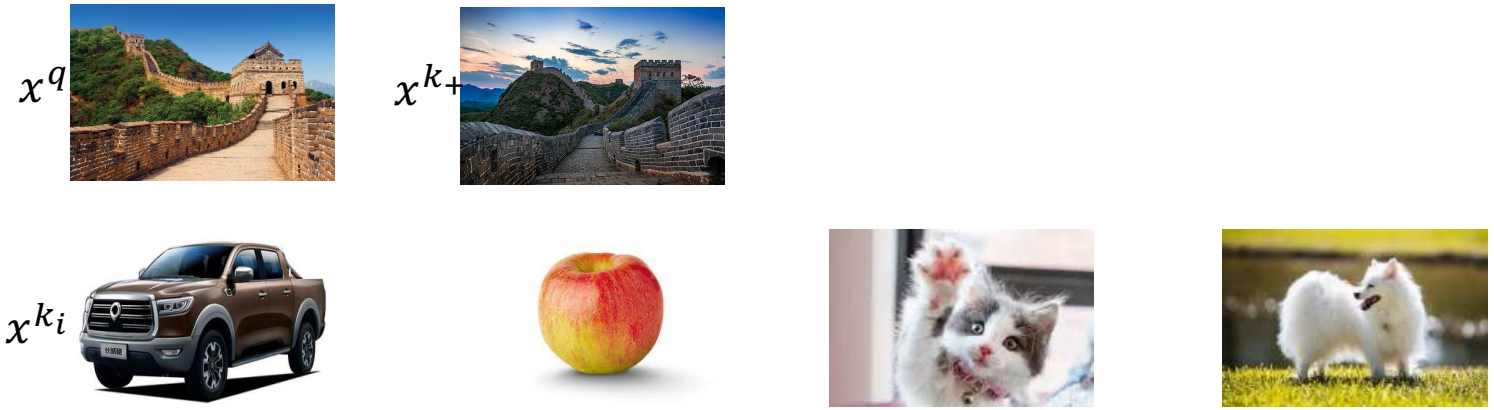
How to align the similar images? How to construct the contrastive pairs?

Self-Supervised Learning

Contrastive Learning



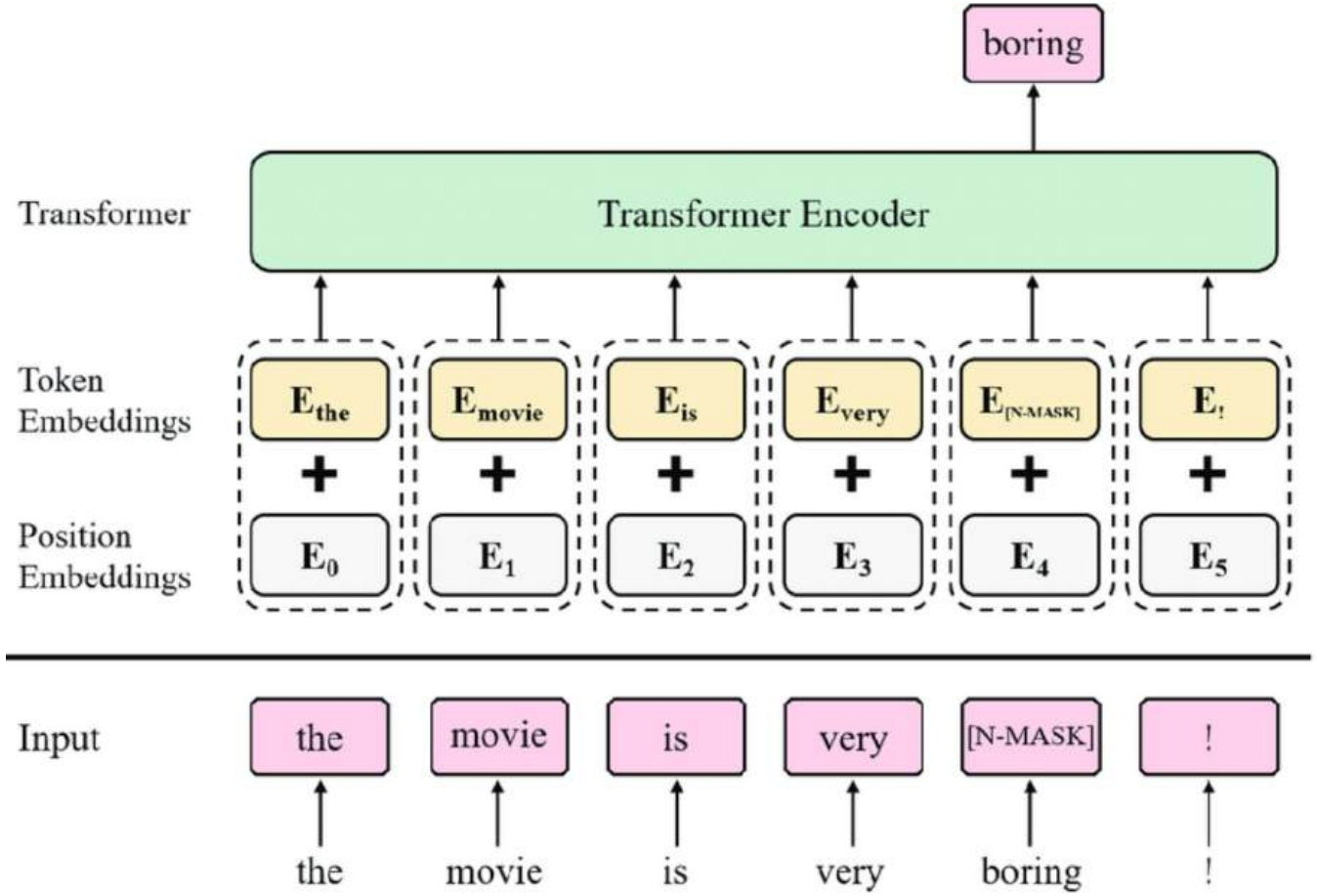
$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$



He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning.", CVPR 2020.

Self-Supervised Learning

Self-supervised learning (Masked Language Modeling) in NLP

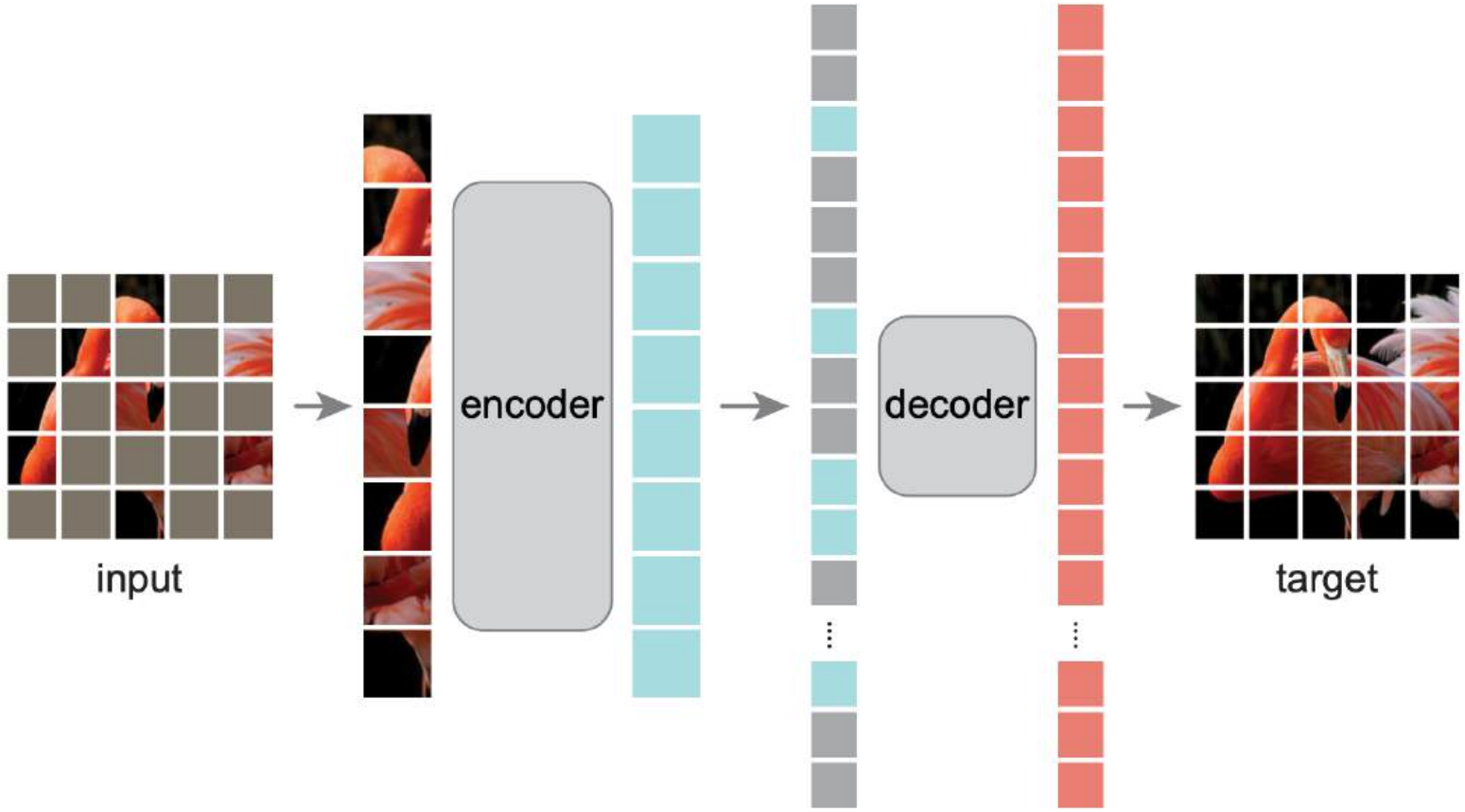


Can such pretext task be transferred from language to vision transformer ?

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding.", NAACL 2019.

Self-Supervised Learning

Masked Image Modeling in Vision (MAE)

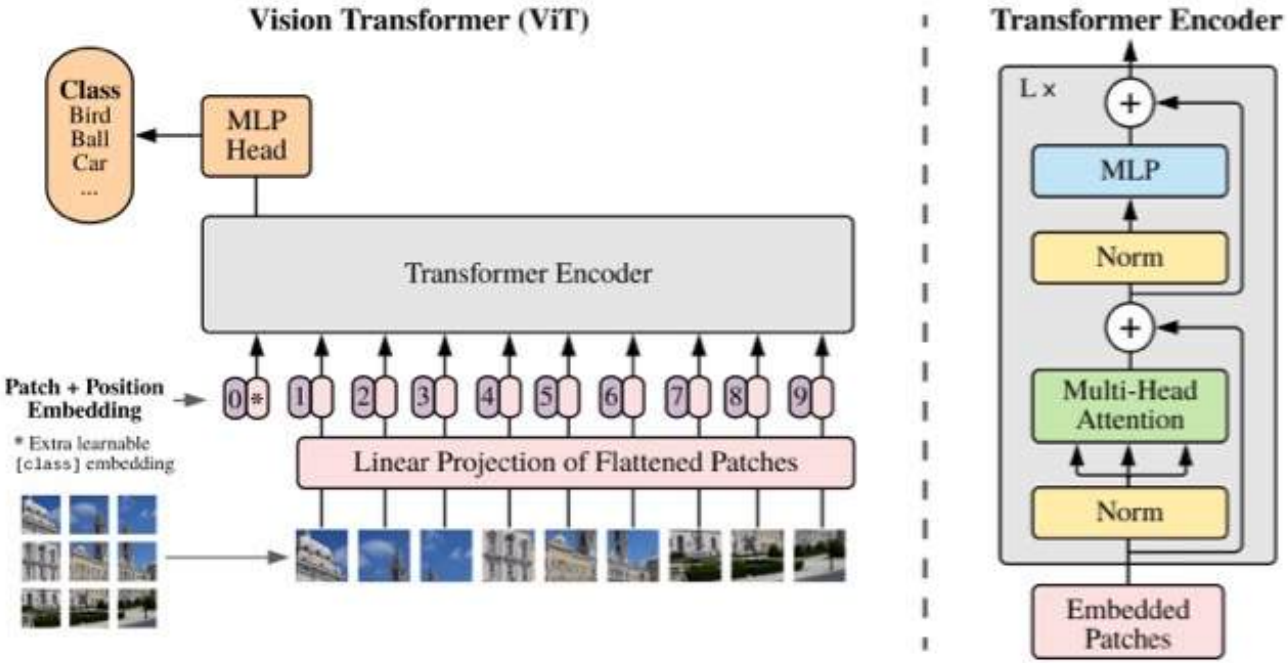


- Different from NLP:
  - The encoder only processes the visible patches, and an extra decoder is used to predict the masked patches
  - Larger masking ratio of 75%, while the default masking ratio in NLP is 15%

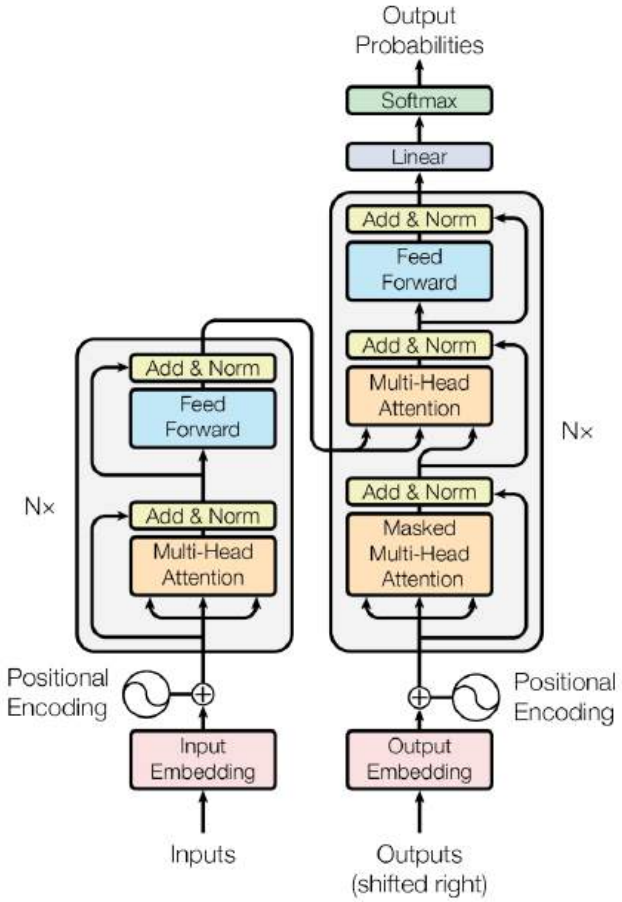
He, Kaiming, et al. "Masked autoencoders are scalable vision learners.", CVPR 2022.

Multi-Modality Training

From the structure perceptive



Vision Transformer Structure

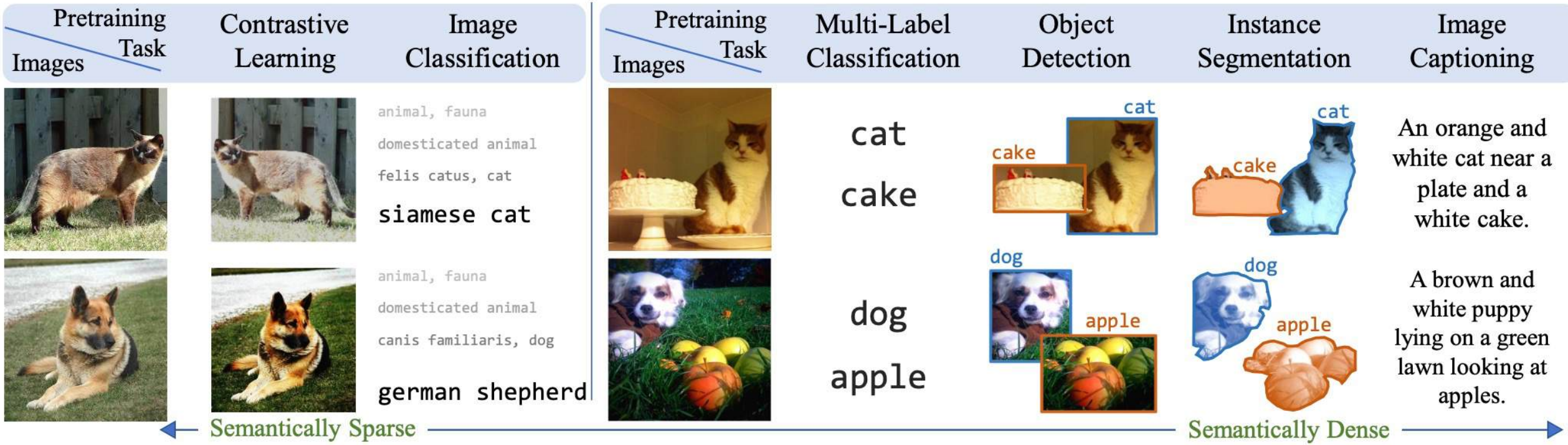


Language Transformer Structure

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." , ICLR 2021.
Vaswani, Ashish, et al. "Attention is all you need." , Neurips 2017.

Multi-Modality Training

The benefits of multi-modality training

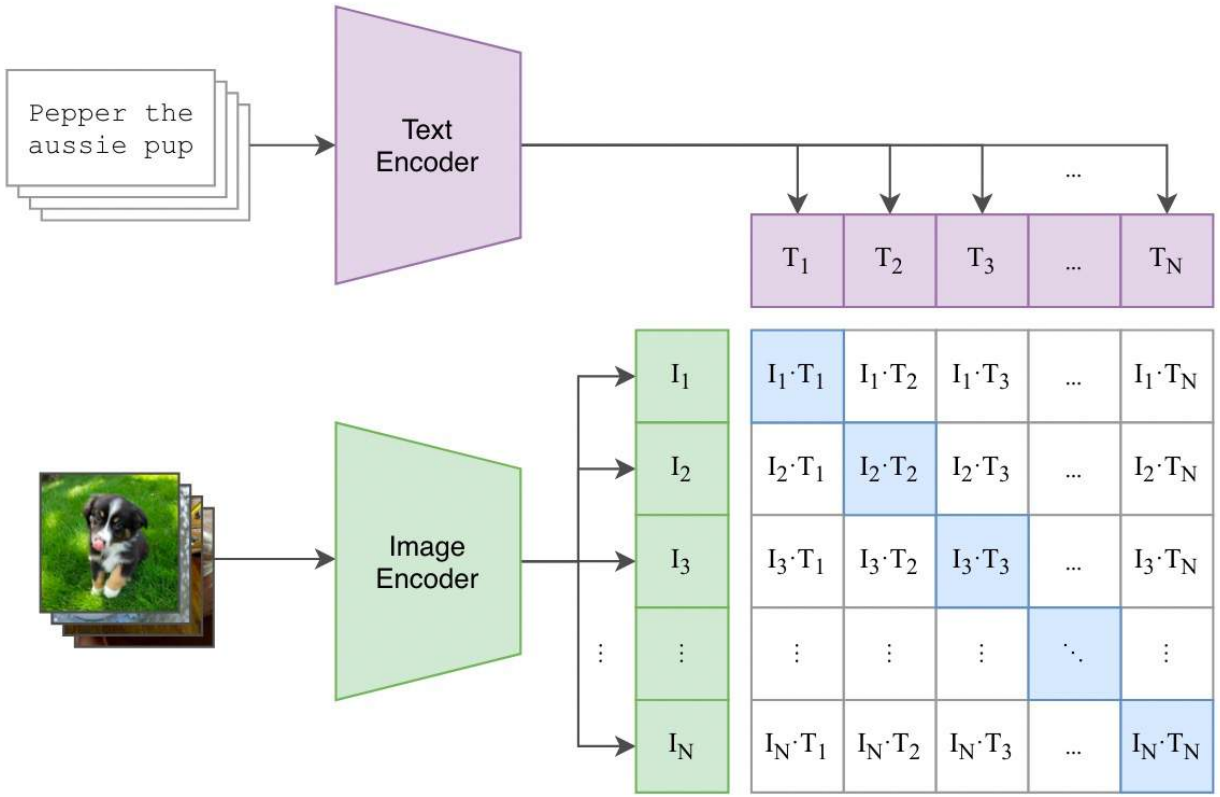


Compared with labels used in image classification, supervisory signals from **language description** contains **dense semantic information**.

Desai, Karan, et al. "Virtex: Learning visual representations from textual annotations.", CVPR 2021.

Multi-Modality Training

Contrastive Pre-training (CLIP)



- Using 400 Million image-text pairs

A group of people on horses on a beach.

• Using 400 Million image-text pairs

A cartoon drawing of sheep watching TV with their babies.

• Using bi-directional contrastive loss, i.e.
image → text $L_{I \to T} = -\log \frac{\exp(I_i * T_i/\tau)}{\sum_k \exp(I_i * T_k/\tau)}$
text → image $L_{T \to I} = -\log \frac{\exp(T_i * I_i/\tau)}{\sum_k \exp(T_i * I_k/\tau)}$

- Using bi-directional contrastive loss, i.e.

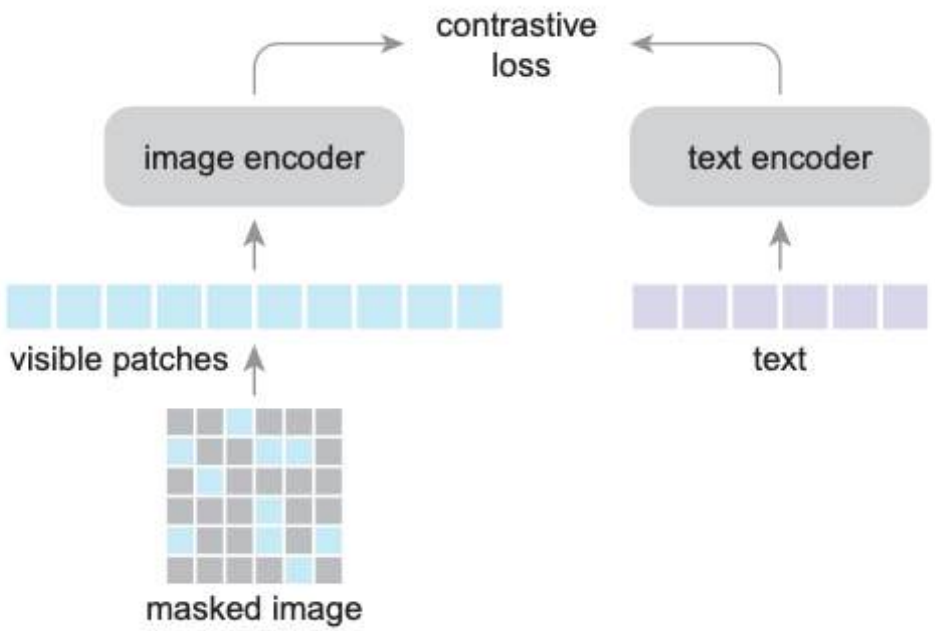
$$\text{image} \to \text{text}\ L_{I \to T} = -\log \frac{\exp(I_i * T_i/\tau)}{\sum_k \exp(I_i * T_k/\tau)}$$

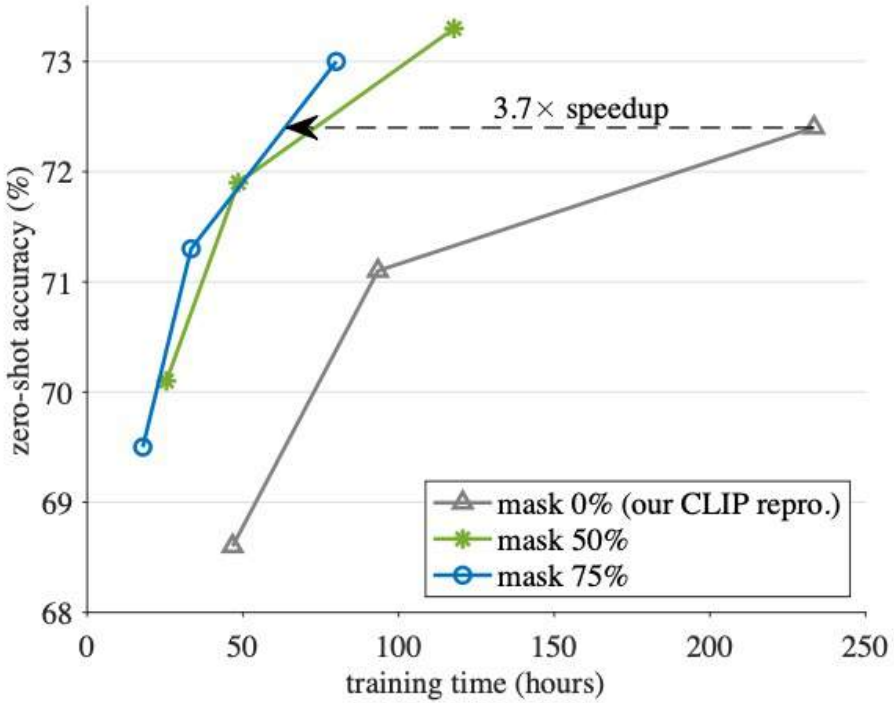
$$\text{text} \to \text{image}\ L_{T \to I} = -\log \frac{\exp(T_i * I_i/\tau)}{\sum_k \exp(T_i * I_k/\tau)}$$

Radford, Alec, et al. "Learning transferable visual models from natural language supervision.", ICML 2021.

Multi-Modality Training

Efficient Contrastive Pre-training (FLIP)



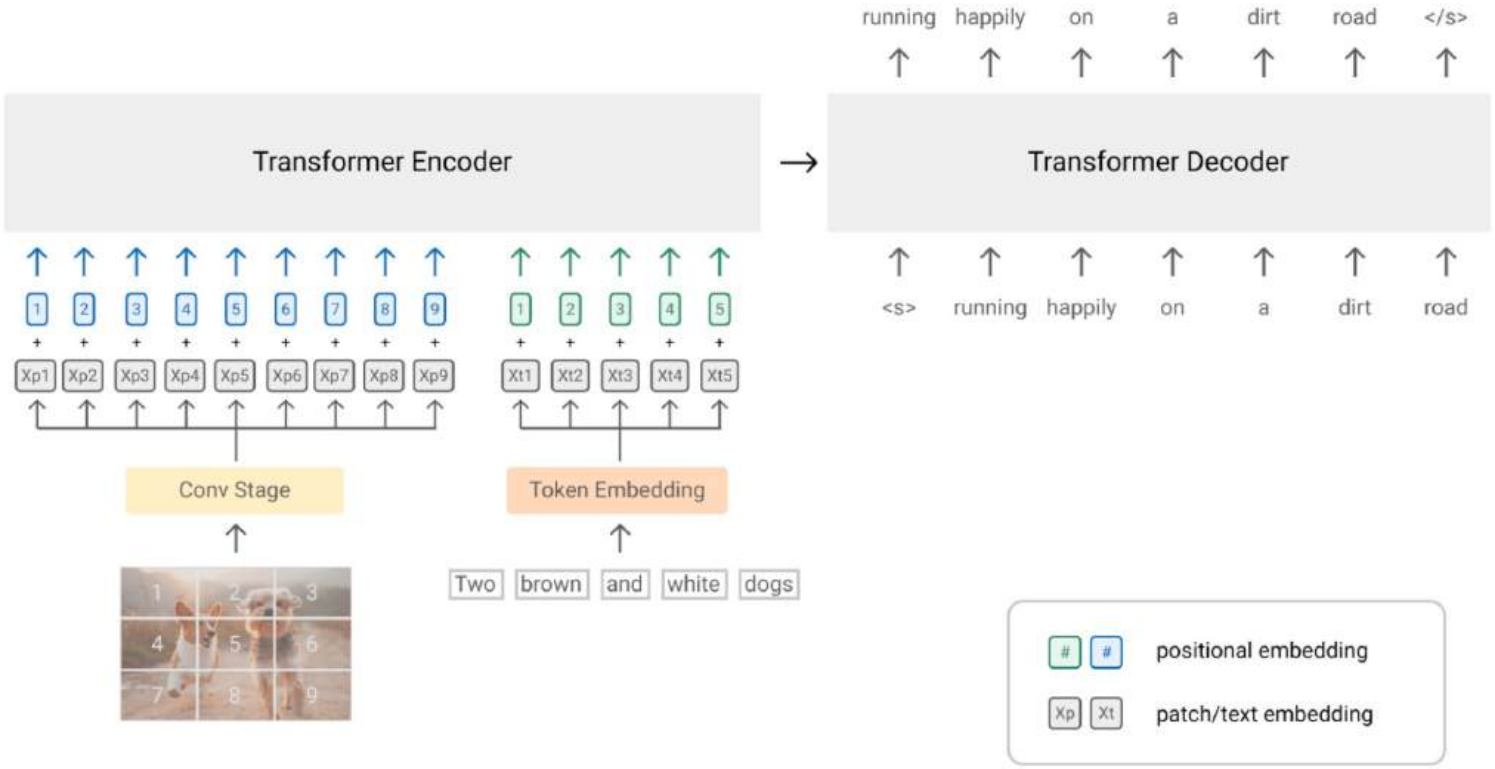


FLIP introduces **image masking** into multi-modality pre-training to improve the training efficiency

Li, Yanghao, et al. "Scaling Language-Image Pre-training via Masking.", CVPR 2023.

Multi-Modality Training

Generative Models (SimVLM)



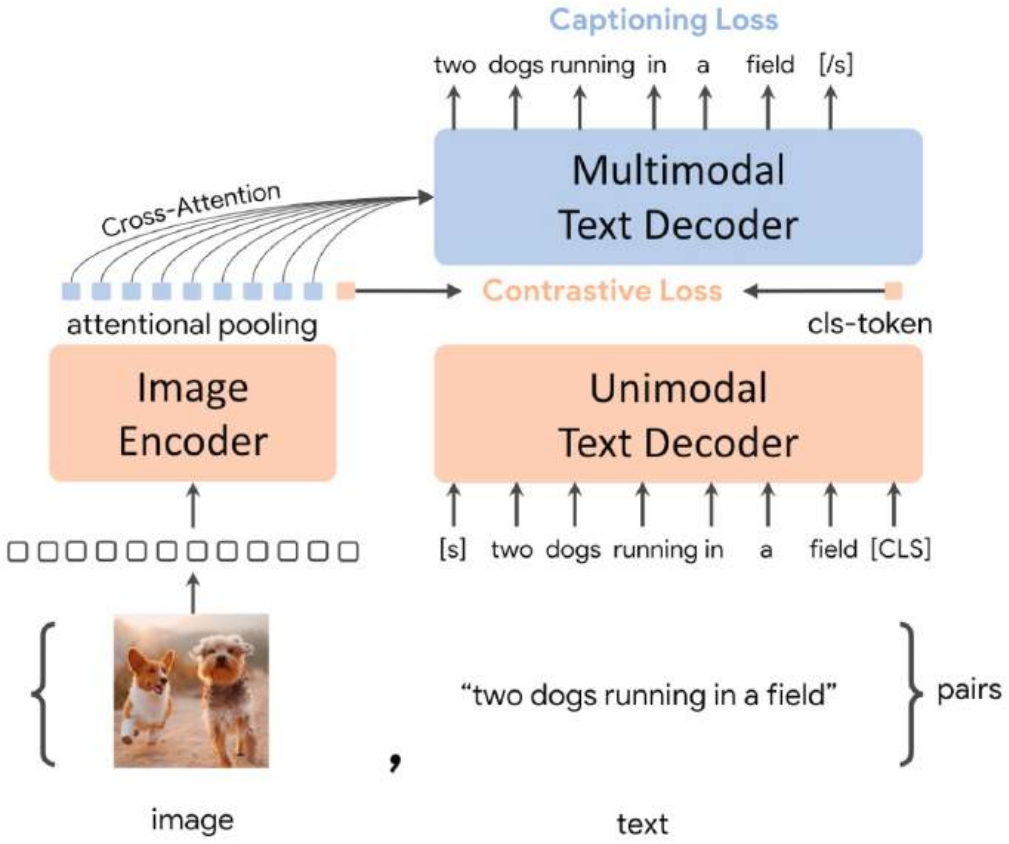
Propose a prefix language modeling task by feeding the **image and former part of the language description as inputs** and asking the model to **predict the left descriptions**

$$\mathcal{L}_{\text{PrefixLM}}(\theta) = -\mathbb{E}_{\mathbf{x}\sim D}\left[\log P_\theta(\mathbf{x}_{\geq T_p}|\mathbf{x}_{<T_p})\right] = -\mathbb{E}_{\mathbf{x}\sim D}\left[\sum_{t=T_p}^{T}\log P_\theta(\mathbf{x}_t|\mathbf{x}_{[T_p,t]},\mathbf{x}_{<T_p})\right]$$

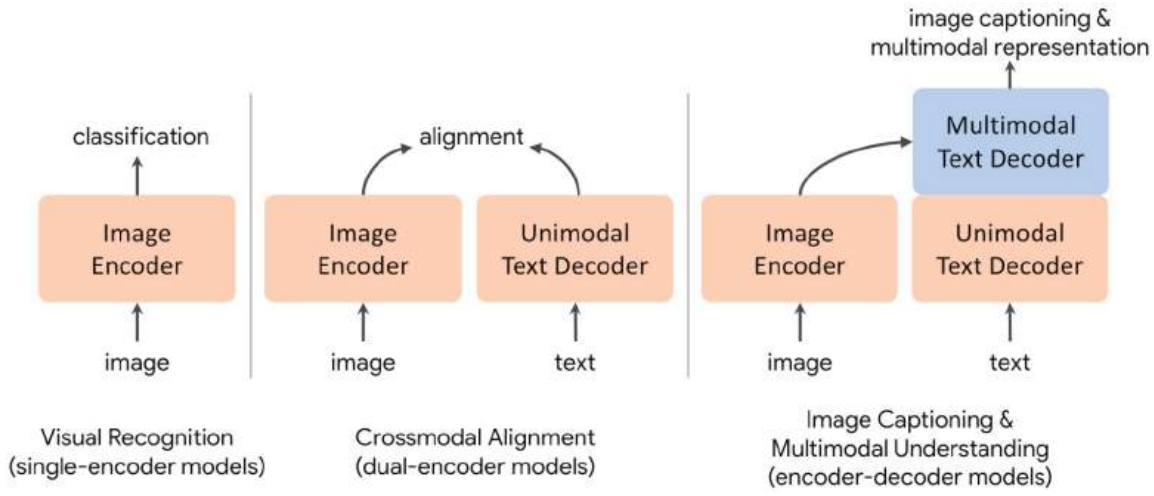
Wang, Zirui, et al. "Simvlm: Simple visual language model pretraining with weak supervision.", ICLR 2022.

Multi-Modality Training

Both Generative and Contrastive Models (CoCa)



CoCa pre-training with contrastive loss and generative loss



Transfer pre-trained CoCa to different tasks

Yu, Jiahui, et al. "Coca: Contrastive captioners are image-text foundation models.", arxiv 2022.

Content

- **Foundation Vision Model**
  - Architecture Design
  - Model Pretraining
  - **Image Generation**
- Applications in Heritage Content Understanding
- Outlook

Image Generation

- Overview

Unconditional Image Generation

Synthesized CelebA-HQ data

Synthesized LSUN-Bedrooms data

Generate images from random noise only.

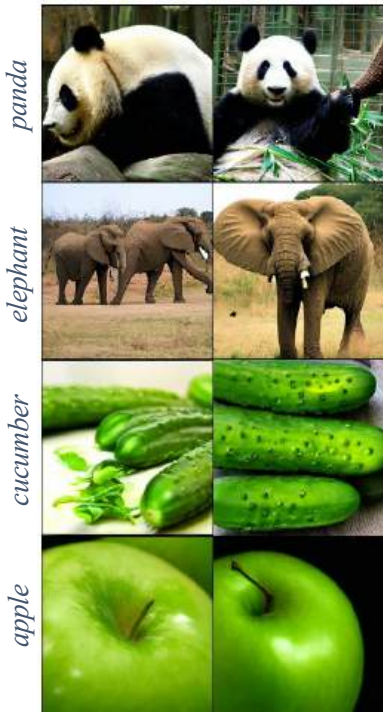
Conditional Image Generation

- e.g., Text-to-Image, Mask-to-Image, …

Text

Semantic Mask

Class Label

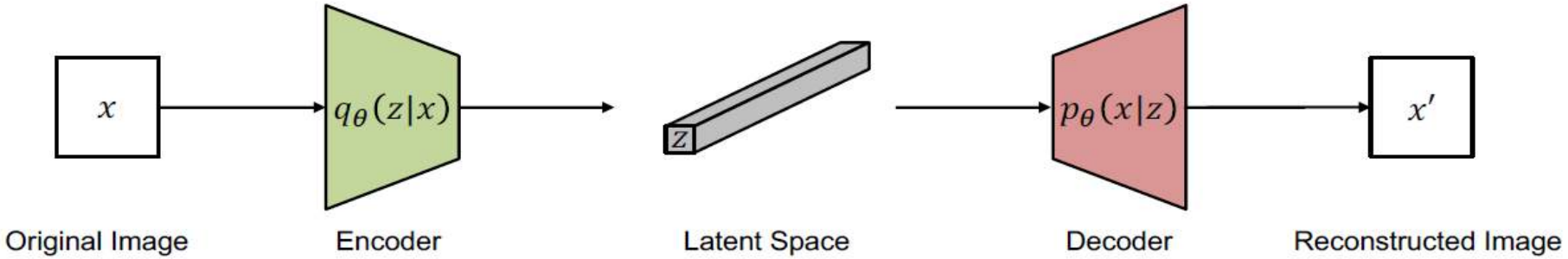


Generate image from random noise and other conditional inputs like text, semantic mask, class label, etc.

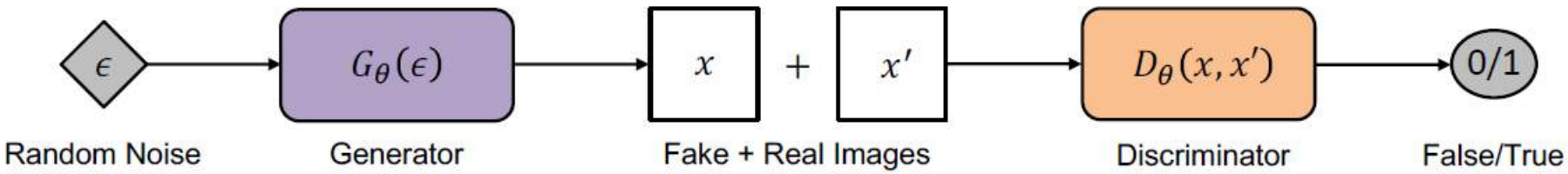
Image Generation

- Method

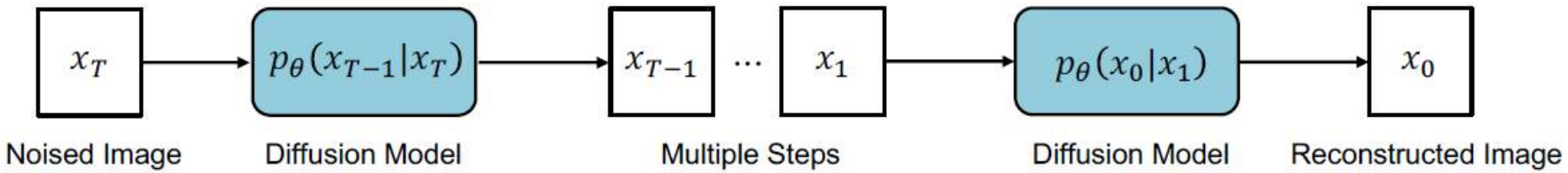
VAE



GAN



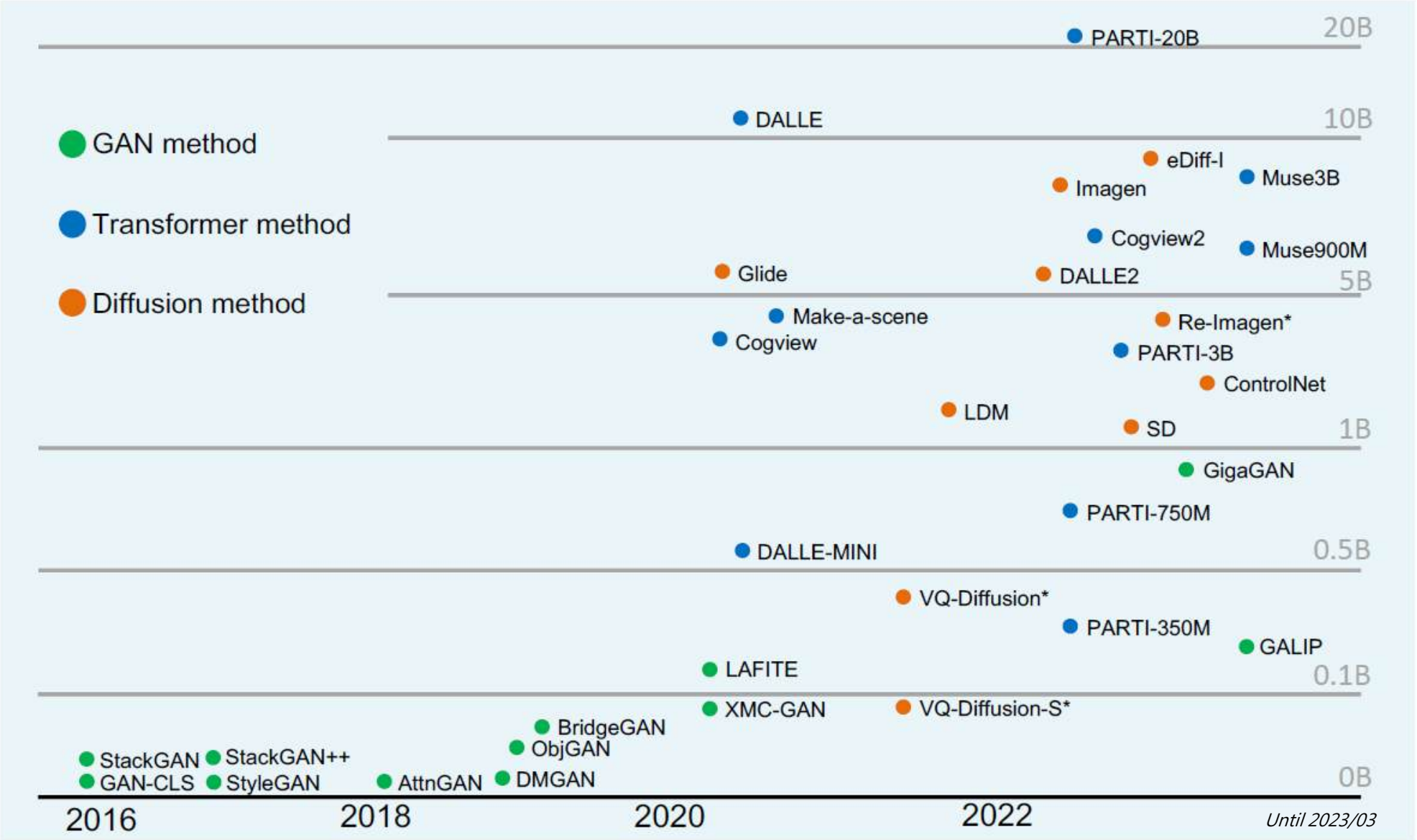
Diffusion Model



Bie, et al: "RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model", CoRR 2309.00810.

Image Generation

- Trend

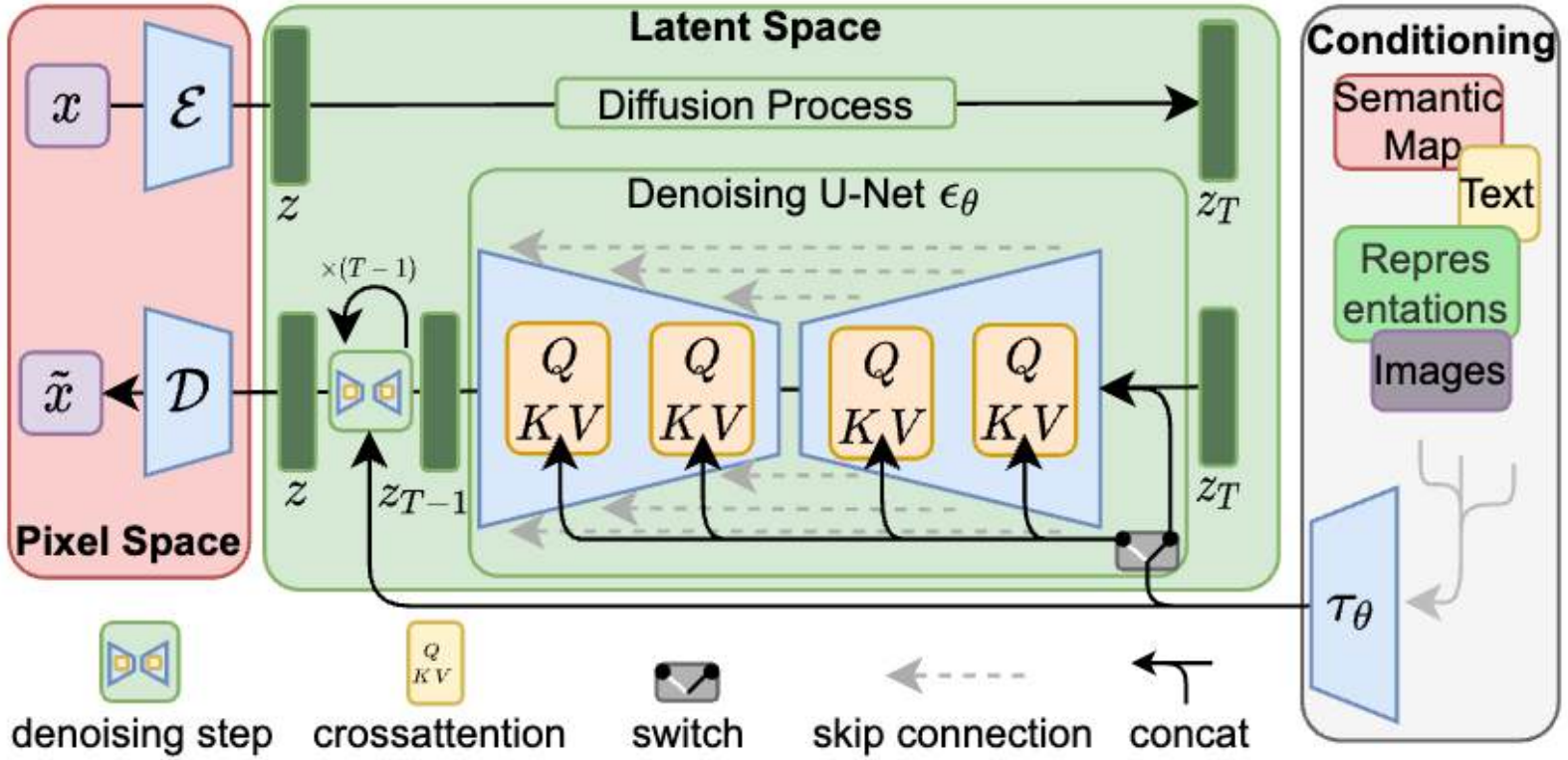


Bie, et al: "RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model", 2023.

Image Generation

- Stable Diffusion

Objective： $L_{LDM} := \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right]$



Condition Injection via cross attention:

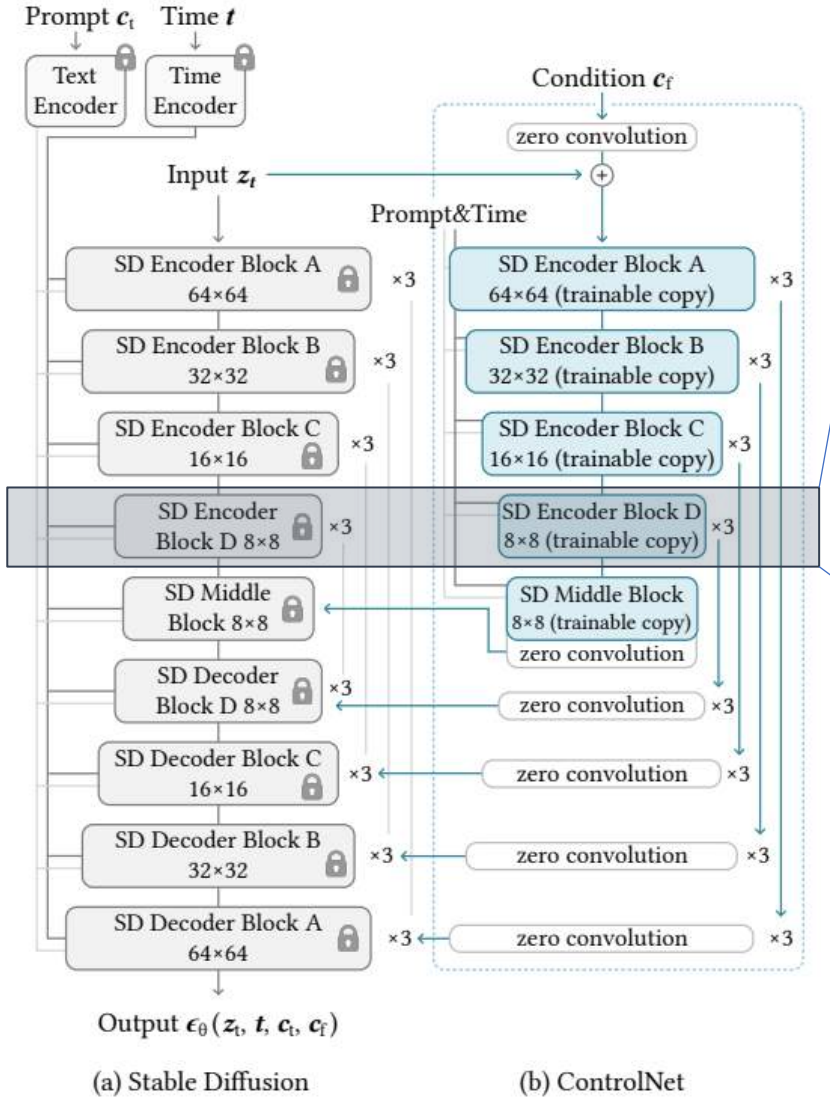
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V,$$

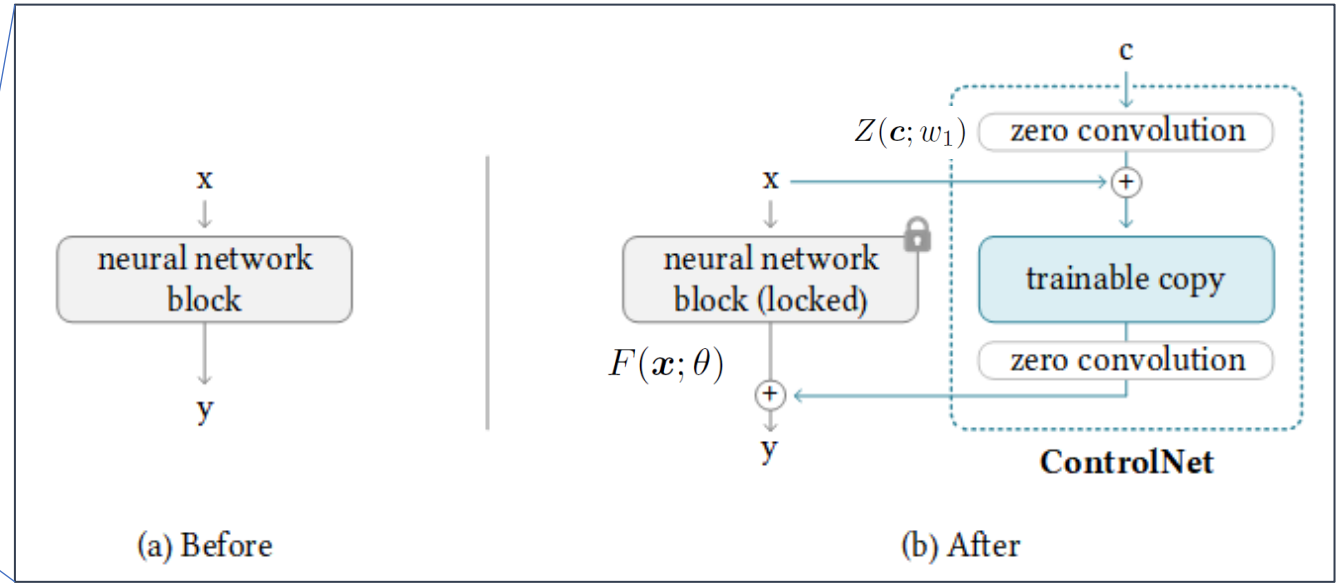
$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y).$$

Rombach, et al: "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR, 2022.

Image Generation

- ControlNet





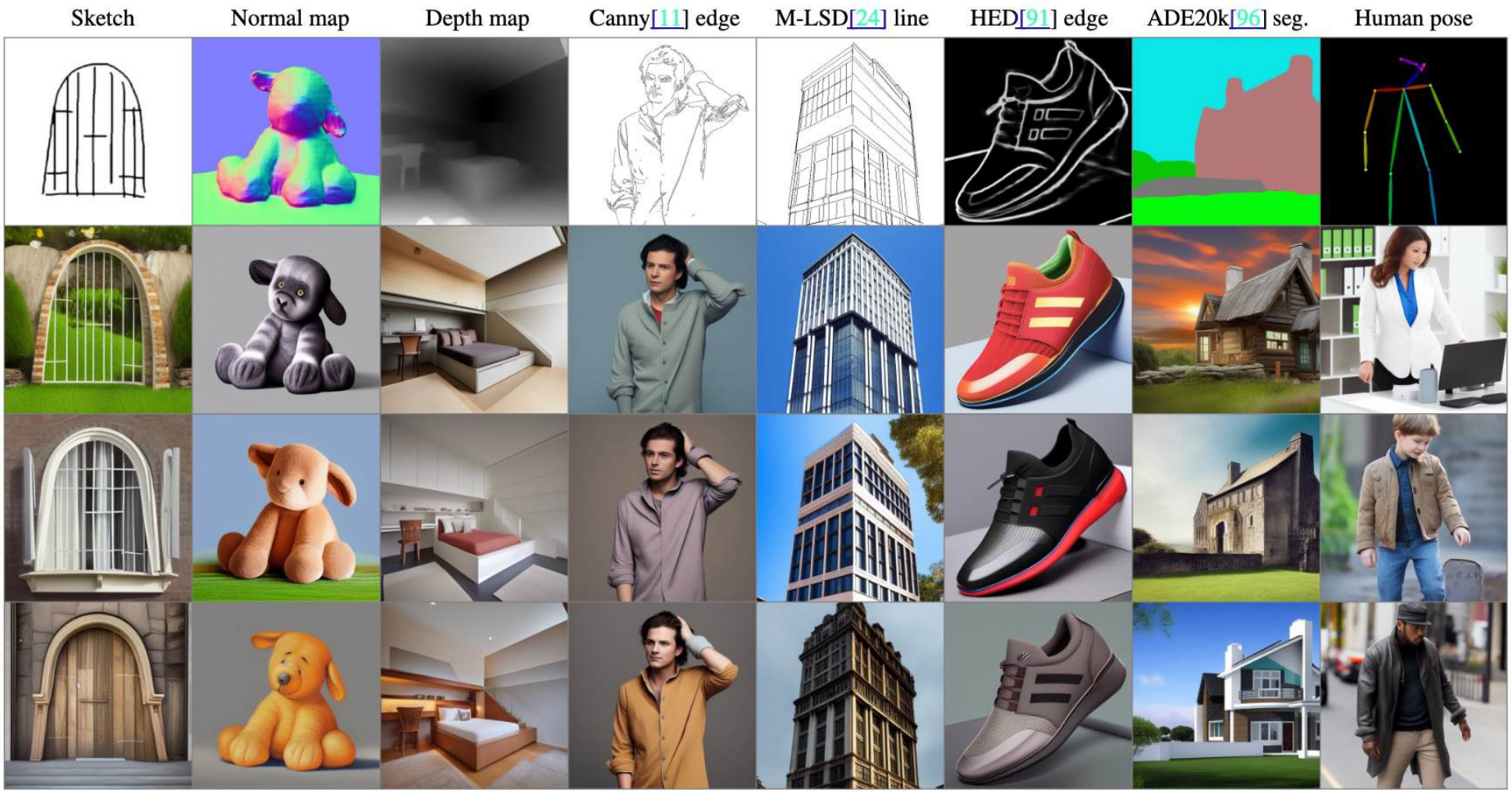
$$y_c = F(x;\theta) + \lambda * Z(F(x + Z(c;w_1);\theta');w_2),$$

Conditioning Scale

Zhang, et al: "Adding Conditional Control to Text-to-Image Diffusion Models", ICCV, 2023.
Qiao, et al: "Learn, Imagine and Create: Text-to-Image Generation from Prior Knowledge", NeurIPS, 2019

Image Generation

- ControlNet

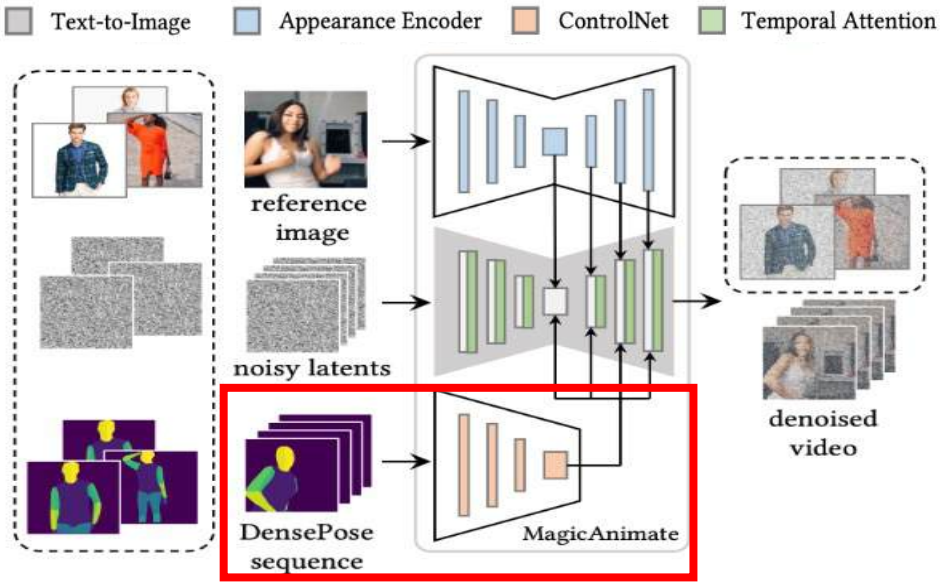


Zhang, et al: "Adding Conditional Control to Text-to-Image Diffusion Models", ICCV, 2023.

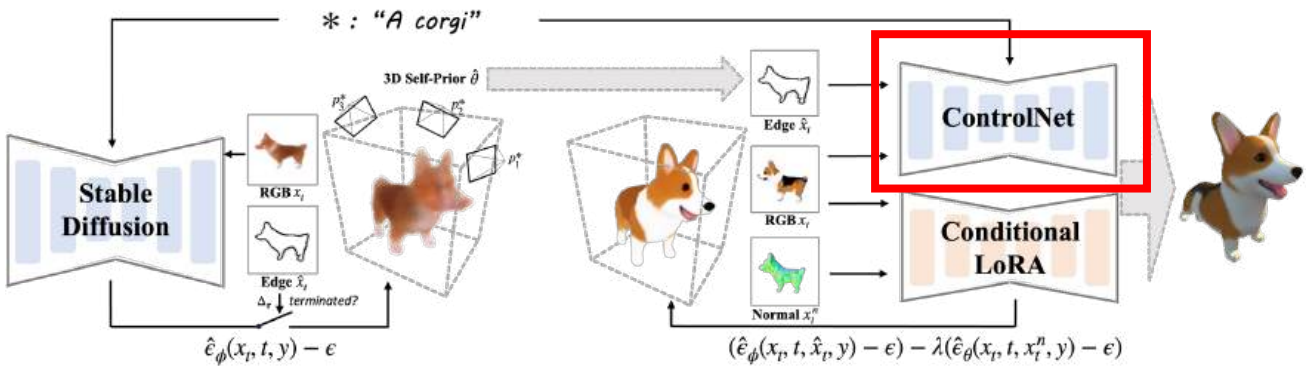
Image Generation

- ControlNet is an influential work for controllable generation.
  - Pose-driven Video Generation
  - Controllable 3D Generation

MagicAnimate



DreamControl



Stage 1: **3D Self-Prior Generation** *Stage 2:* **Control-Based Score Distillation**

Elon musk, using a laptop

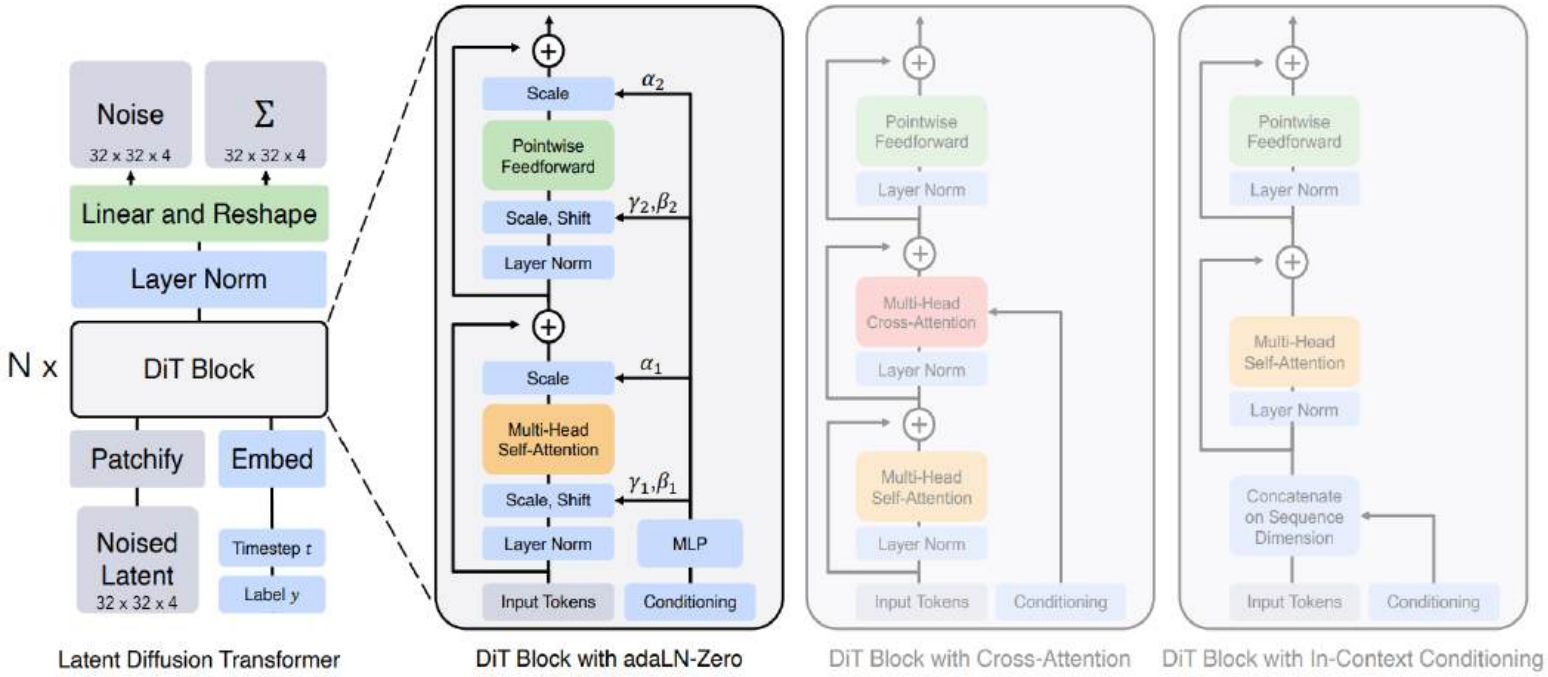
Batman is riding a moto

An astronaut is riding a horse

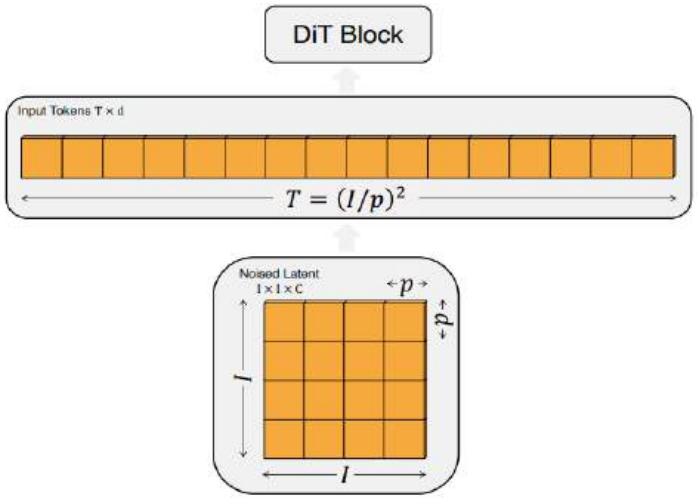
Chen, et al: "DreamControl: Control-Based Text-to-3D Generation with 3D Self-Prior", CVPR, 2024.
Xu, et al: "MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model", CVPR, 2024.

Image Generation

- Diffusion Transformer (DiT)
  - Many designs are borrowed from ViT.
  - Adaptive layer norm works best in incorporating conditioning compared with cross-attention and extra input tokens.



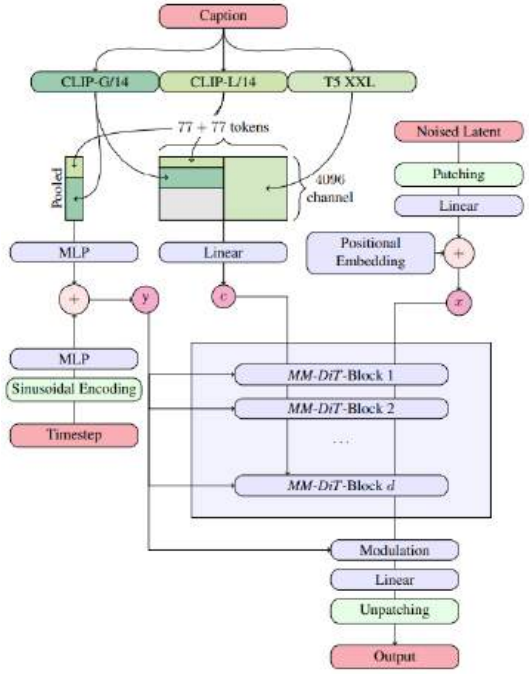
Input specifications for DiT



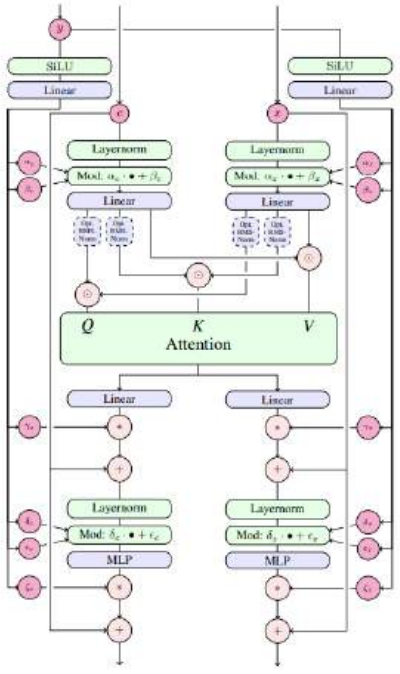
William et al: "Scalable Diffusion Models with Transformers", ICCV 2023

Image Generation

- Applications of Diffusion Transformer (DiT)
  - **Stable Diffusion 3** (Image Generation)
    - A diffusion transformer architecture (MM-DiT)
    - Flow matching

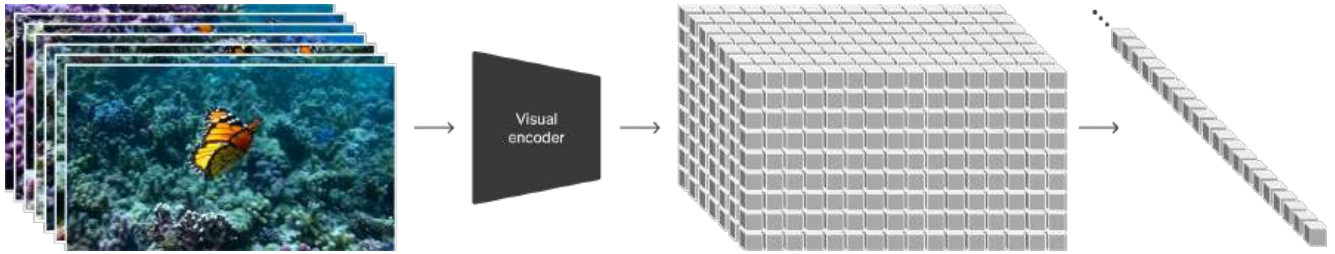


(a) Overview of all components.



(b) One *MM-DiT* block

  - **OpenAI Sora** (Video generation as world simulators)
    - Spacetime visual patches
    - Transformer-based diffusion models



An image of a realistic cloud that spells "SORA".

In an ornate, historical hall, a massive tidal wave peaks and begins to crash. Two surfers, seizing the moment, skillfully navigate the face of the wave.

OpenAI Sora: https://openai.com/index/sora/

Patrick et al: "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis"

Content

- Foundation Vision Model
  - Architecture Design
  - Model Pretraining
  - Image Generation

- **Applications in Heritage Content Understanding**

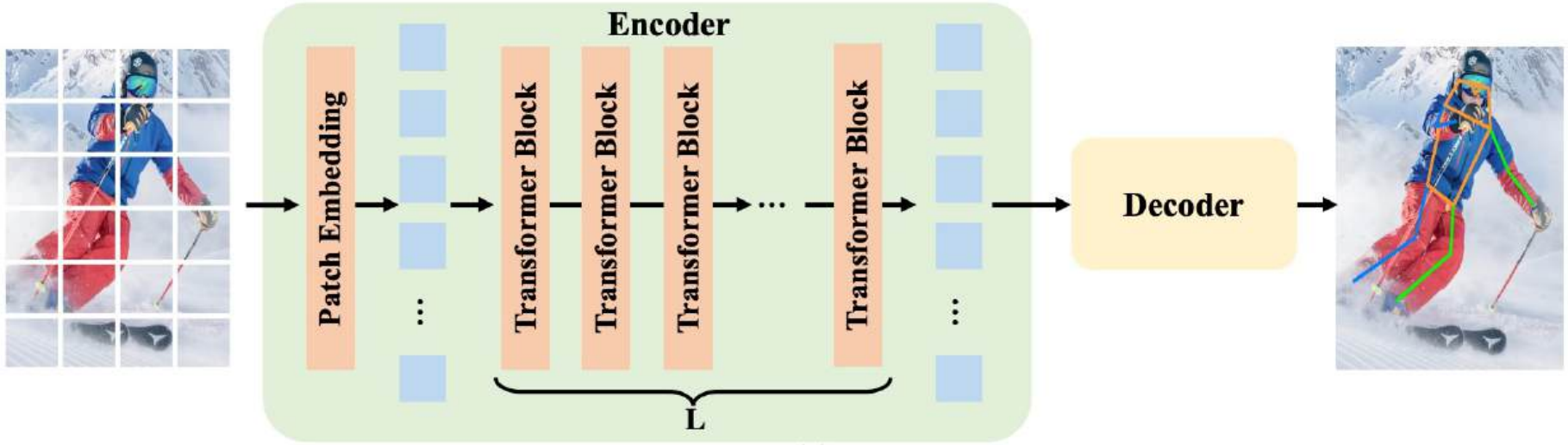
- Outlook

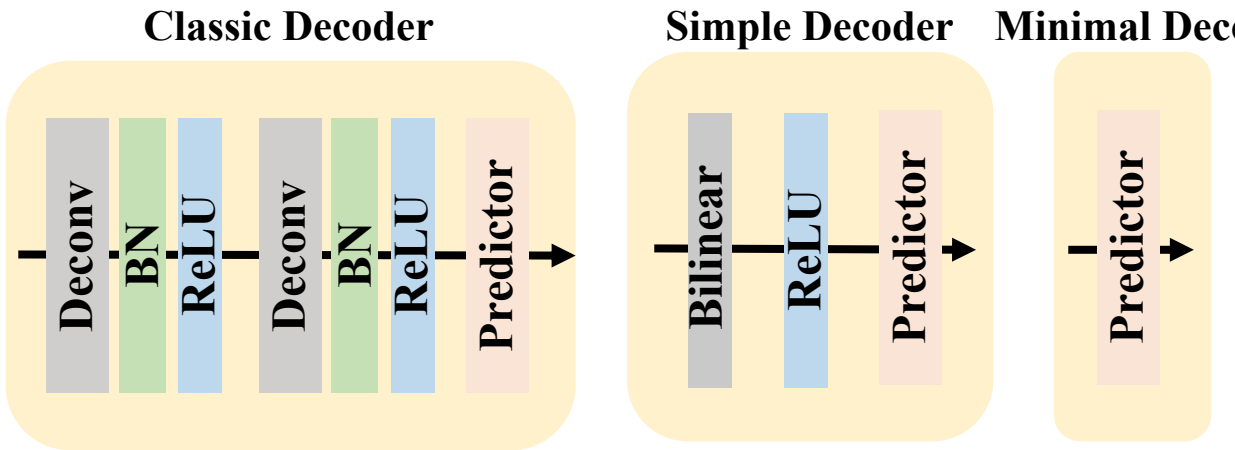
Heritage Content Understanding and Generation

- Visual Understanding
  - Image Classification
  - Object Detection
  - Instance Segmentation
  - Pose Estimation
  - Text Spotting
- Cross-modal Understanding
  - Visual Grounding
  - Visual Question Answering
- Visual Content Generation
  - Image Generation

LVMs for Different Downstream Tasks

Human Pose Estimation: ViTPose & ViTPose++



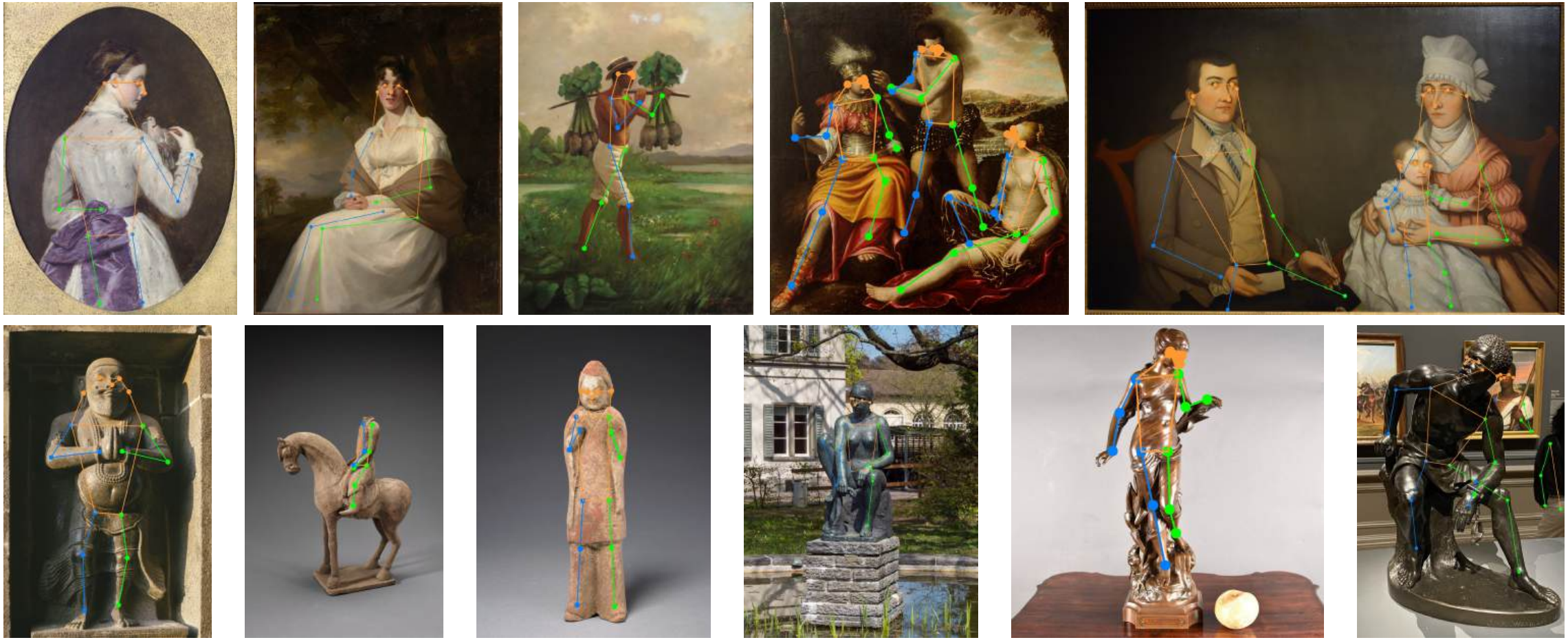


- Vision Transformer (e.g., ViT or ViTAE) as encoder
- Scaling up to 1 billion parameters
- The decoder could be as simple as a single (1*1) convolution layer
- Solve multiple pose estimation tasks simultaneously

Xu, Yufei, et al. "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation." NeurIPS 2022.
Xu, Yufei, et al. " ViTPose++: Vision Transformer for Generic Body Pose Estimation", IEEE TPAMI 2023.

LVMs for Different Downstream Tasks

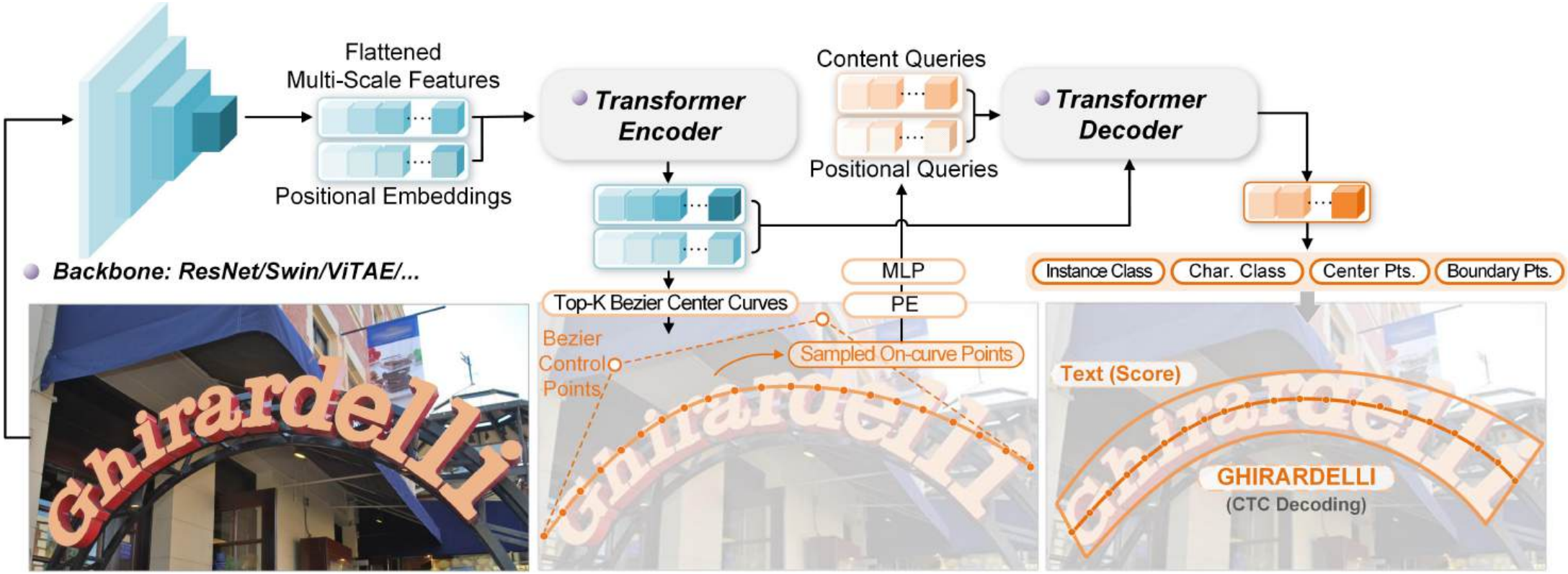
Human Pose Estimation: ViTPose & ViTPose++



Xu, Yufei, et al. "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation." NeurIPS 2022.
Xu, Yufei, et al. " ViTPose++: Vision Transformer for Generic Body Pose Estimation", IEEE TPAMI 2023.

LVMs for Different Downstream Tasks

Text Spotting: DeepSolo & DeepSolo++



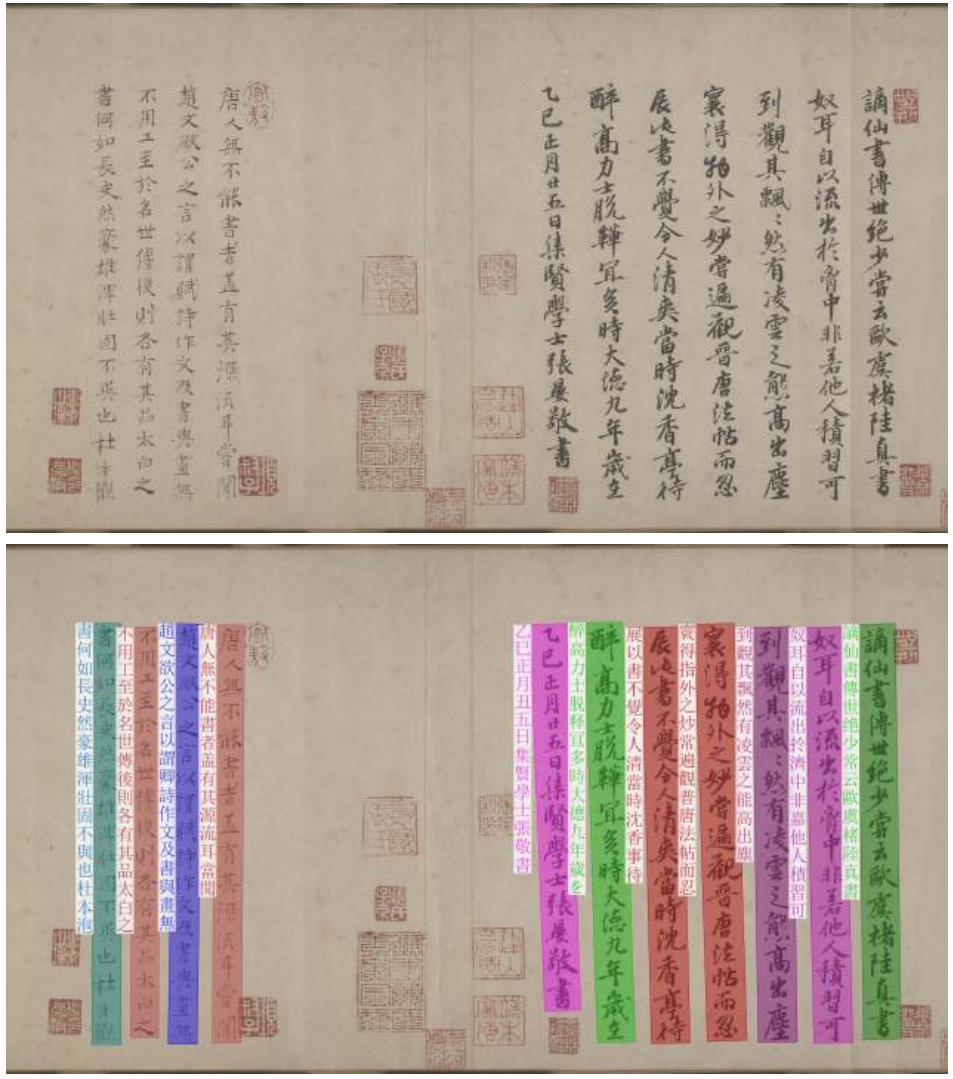
- Vision Transformer as encoder
- Learnable explicit point queries
  - Bezier curve proposal and on-curve point sampling
- A single transformer decoder for multiple tasks

- **Simplicity** of structure and training
- **Efficiency** of training and inference
- **Extensibility** of character class, language, and task

Ye, Maoyuan, et al. "DeepSolo: Let Transformer Decoder with Explicit Points Solo for Text Spotting", CVPR 2023.
Ye, Maoyuan, et al. "DeepSolo++: Let Transformer Decoder with Explicit Points Solo for Multilingual Text Spotting", CoRR 2305.19957.

LVMs for Different Downstream Tasks

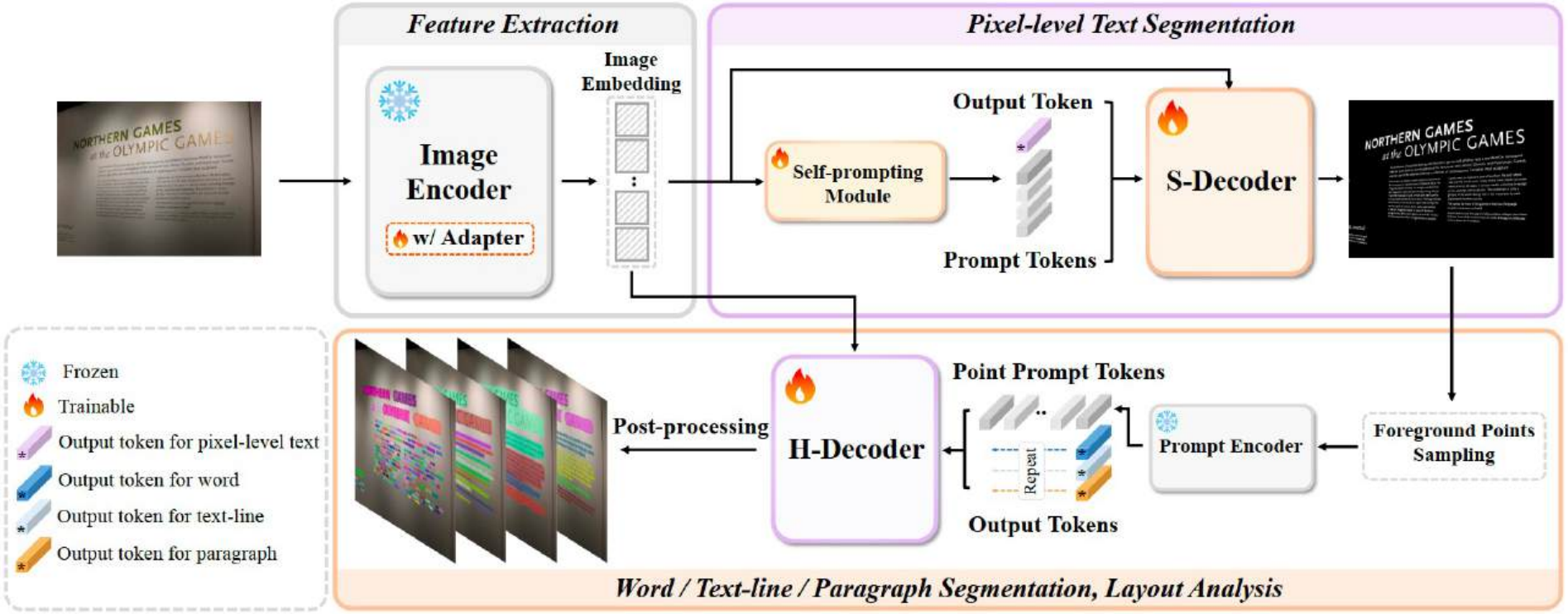
Text Spotting: DeepSolo & DeepSolo++



Ye, Maoyuan, et al. "DeepSolo: Let Transformer Decoder with Explicit Points Solo for Text Spotting", CVPR 2023.
Ye, Maoyuan, et al. "DeepSolo++: Let Transformer Decoder with Explicit Points Solo for Multilingual Text Spotting", CoRR 2305.19957.

LVMs for Different Downstream Tasks

Text Segmentation: Hi-SAM



- Vision Transformer as encoder
- Adapt SAM's general segmentation capability to the text image domain
- The first study of text segmentation across four hierarchies

Ye, Maoyuan, et al. "Hi-SAM: Marrying Segment Anything Model for Hierarchical Text Segmentation", IEEE TPAMI Revision 2024.

LVMs for Different Downstream Tasks

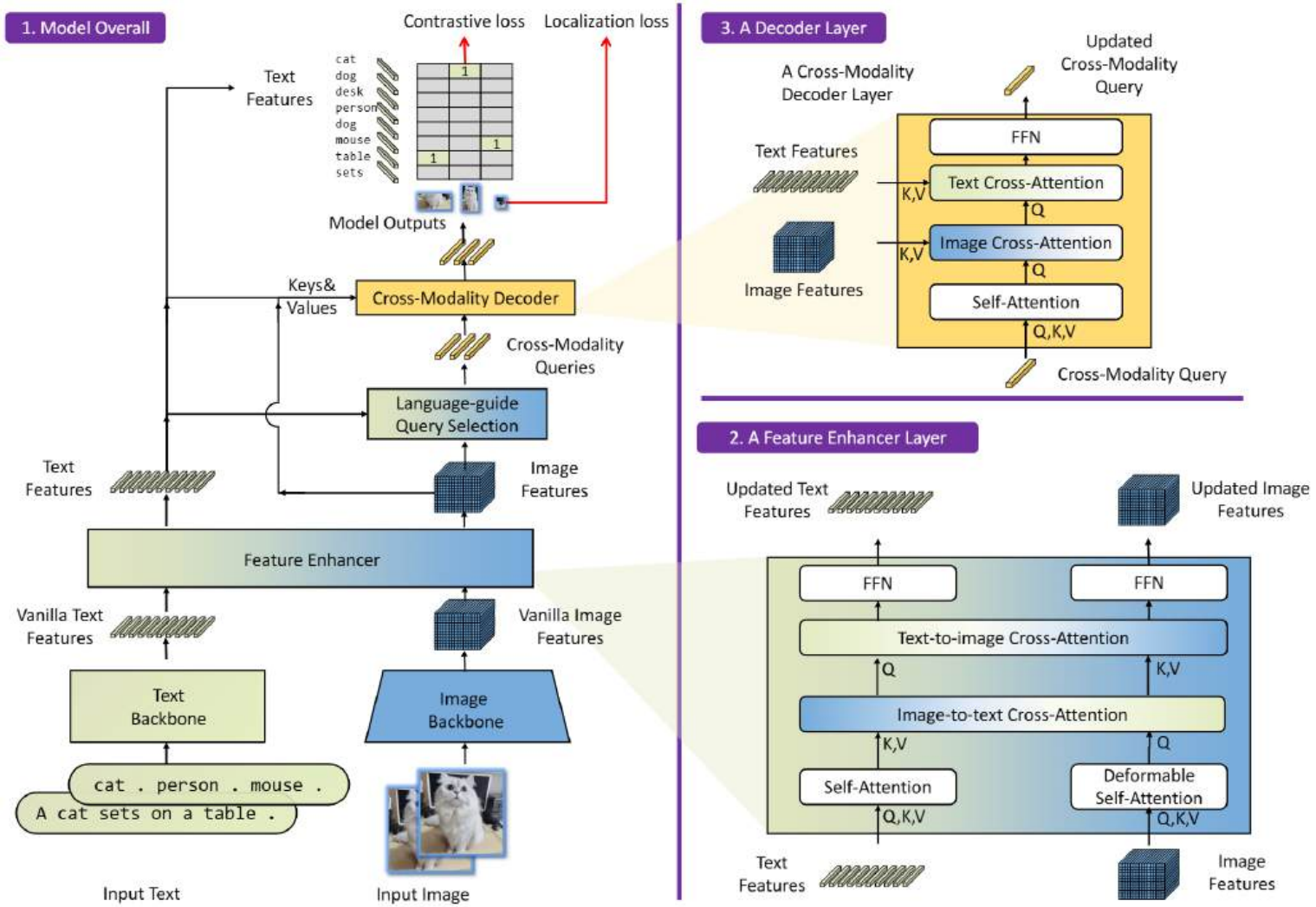
Text Segmentation: Hi-SAM

Cultural Relic Images | Text Character Seg. | Word Segmentation | Line Segmentation | Paragraph Segmentation

Ye, Maoyuan, et al. "Hi-SAM: Marrying Segment Anything Model for Hierarchical Text Segmentation", IEEE TPAMI Revision 2024.

LVMs for Different Downstream Tasks

Visual Grounding: Grounding DINO



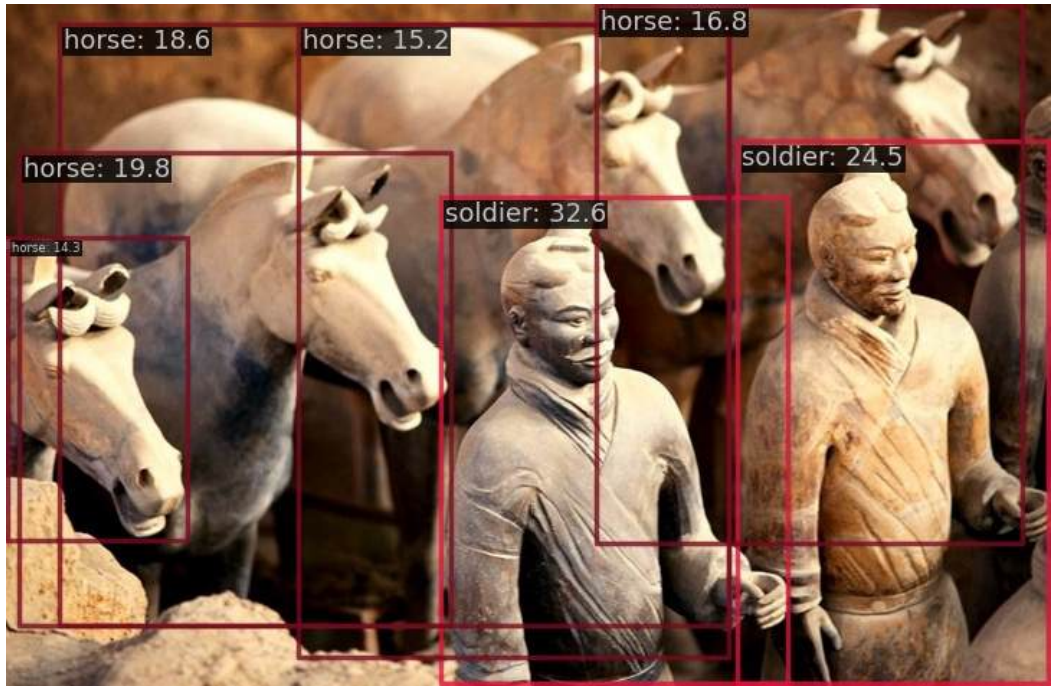
- Transformer-Based Architecture
  - Query-Based Object Detection
  - Cross-Modality Decoder
  - Transformer-based Feature Enhancer
- Grounding via Vision-Language Pretraining
  - Introducing language to a closed-set detector for open-set concept generalization
  - Multimodal Feature Fusion
  - Visual-Text Contrastive Learning

Liu, Shilong, et al. "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection", ECCV 2024.

LVMs for Different Downstream Tasks

Visual Grounding: Grounding DINO

Description: The image shows a person riding a horse and a human head with a lion's body.

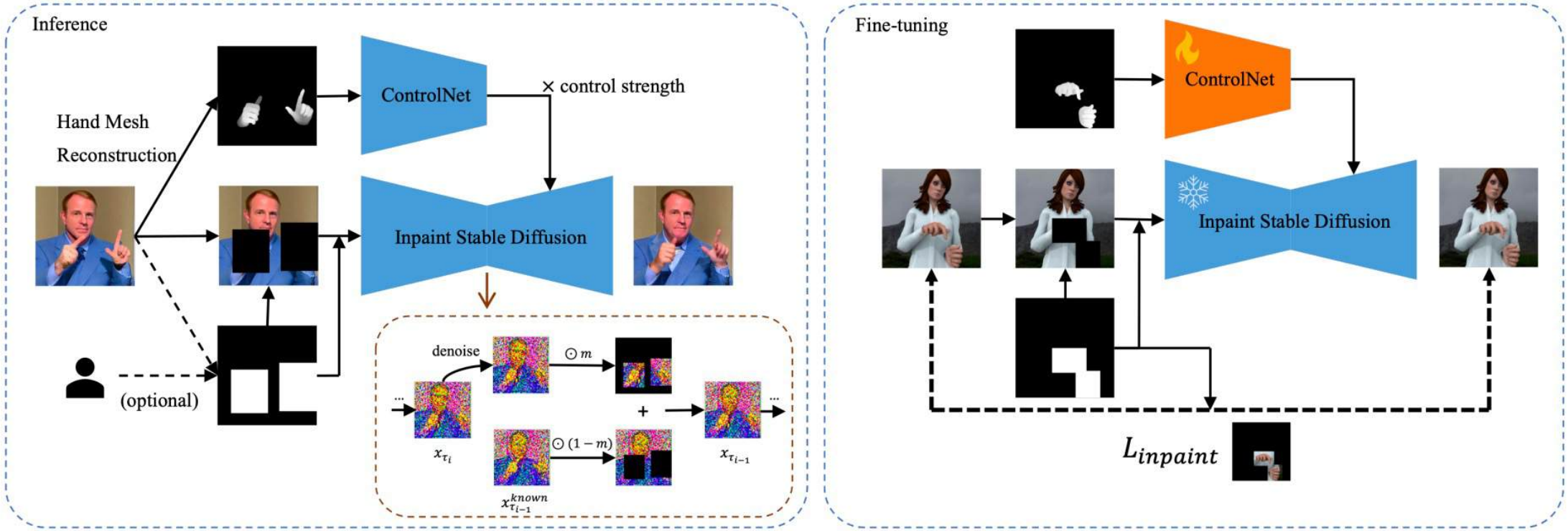


Description: The image depicts numerous ancient Chinese soldiers and horses

Liu, Shilong, et al. "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection", ECCV 2024.

LVMs for Different Downstream Tasks

Image Generation: HandRefiner



- Built upon Stable Diffusion and ControlNet
- Refining malformed hands in generated images as conditional inpainting
- Leverage hand shape prior as condition
- Leverage synthetic data for training

Lu, Wenquan, et al. "HandRefiner: Refining Malformed Hands in Generated Images by Diffusion-based Conditional Inpainting", ACM Multimedia 2024.

LVMs for Different Downstream Tasks

Image Generation: HandRefiner

Lu, Wenquan, et al. "HandRefiner: Refining Malformed Hands in Generated Images by Diffusion-based Conditional Inpainting", ACM Multimedia 2024.

Content

- Foundation Vision Model
  - Architecture Design
  - Model Pretraining
  - Image Generation
- Applications in Heritage Content Understanding
- **Outlook**

LVMs for Heritage Content Understanding

Deciphering Ancient Languages

LVMs for Heritage Content Understanding

World Heritage Preservation

Galapagos Islands (Ecuador)

Pyramids of Cairo (Egypt)

Great Wall (China)

https://heritage.earsel.org/
https://masonrymagazine.com/Default?pageID=14191
https://intelligence.airbus.com/newsroom/satellite-image-gallery/pleiades-neo/cairo-egypt-very-high-resolution-imagery/

LVMs for Heritage Content Understanding

Creating Engaging, Interactive Experiences

https://youtu.be/H61_8MaC5C8

Thank you!