Transformers in Vision: A Survey

Phitchaya Faramnuayphol

Outline

- Transformers for Object Detection
- Transformers for Segmentation
- Transformers for Image and Scene Generation
- Transformers for Low-level Vision
- Transformers for Multi-Modal Tasks
- Transformers for 3D Analysis
- Open Challenges

Transformers for Object Detection

Detection Transformers with CNN Backbone

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
Object Detection	DETR [13]	Linear projection layer to reduce CNN feature dimension, Spatial positional embedding added to each multi-head self-attention layer of both encoder and decoder. Object queries (output positional encoding) added to each multi-head self-attention layer of decoder.	2D Image	Class labels	Hungarian loss based on bipartite matching between predicted and ground truths
	D-DETR [14]	Deformable Transformer consists of deformable attention layers to introduce sparse priors in Transformers, Multi-scale attention module.	2D Image	Class labels	Hungarian loss

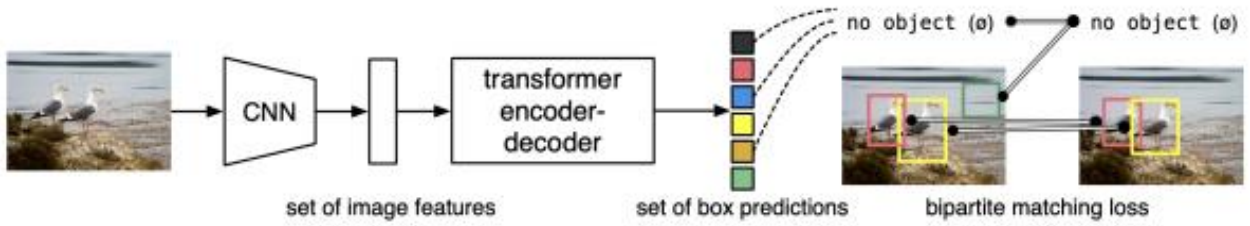


Fig. 7: Detection Transformer (DETR) [13] treats the object detection task as a set prediction problem and uses the Transformer network to encode relationships between set elements. A bipartite set loss is used to uniquely match the box predictions with the ground-truth boxes (shown on the *right* two columns). In case of no match, a *'no object'* class prediction is selected. Its simple design with minimal problem-specific modifications can beat a carefully built and popular Faster R-CNN model. Figure from [13].

Transformers for Object Detection

Task	Method	Metric	Dataset	Performance	Highlights	Limitations
Object Detection	DETR [13] ECCV'20	AP	COCO	44.9	**a)** Use of Transformer allows end-to-end training pipeline for object detection, **b)** Removes the need for hand-crafted post-processing steps.	**a)** Performs poorly on small objects, **b)** Requires long training time to converge.
	D-DETR [14] ICLR'21	AP	COCO	43.8	**a)** Achieves better performance on small objects than DETR [13], **b)** Faster convergence than DETR [13]	Obtain SOTA results with 52.3 AP but with two stage detector design and test time augmentations.

Transformers for Segmentation

- Cross-model Self-attention (CMSA)

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
Referring Image	CMSA [15]	Multimodal feature, Cross-modal self-attention on multiple levels and their fu-	2D Image + Language expression	Segmentation mask	Binary cross-entropy loss

- Panoptic segmentation

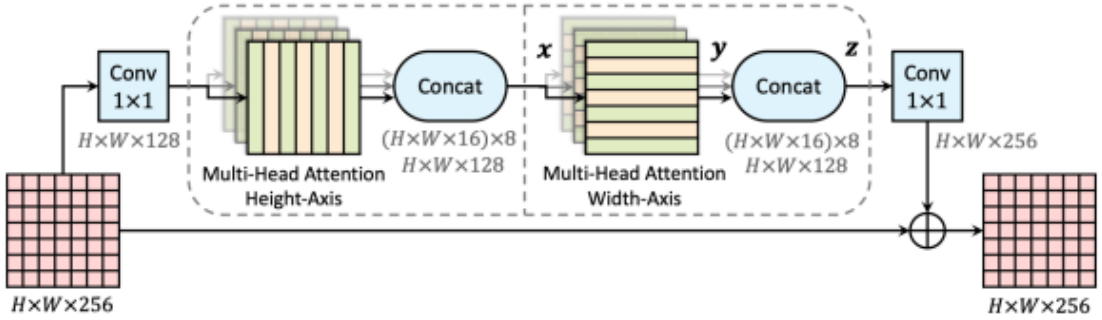


Fig. 8: Axial attention module [133] that sequentially applies multi-head axial attention operations along height and width axes. Image from [133].

Transformers for Image and Scene Generation

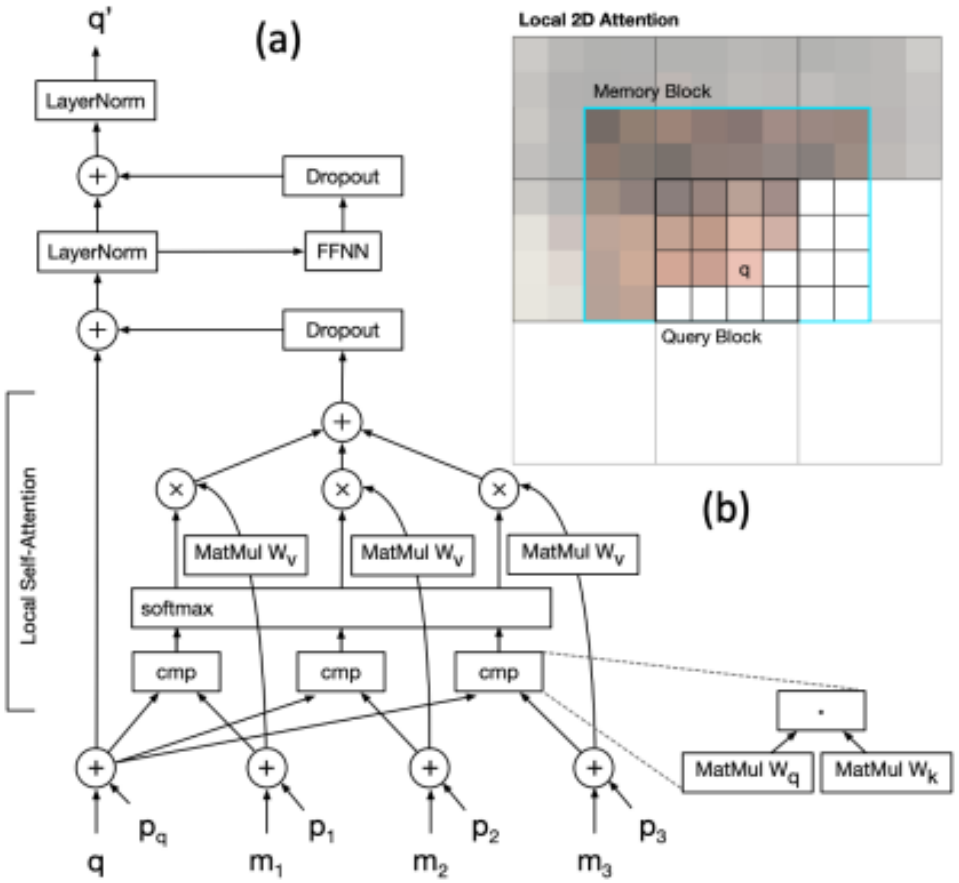
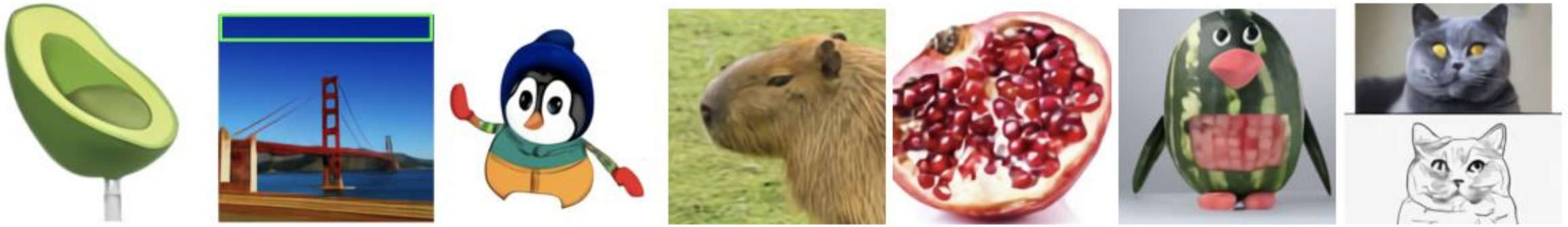


Fig. 9: (a) Self-attention block in Image Transformer [142]. Given one channel for a pixel q, the block attends to the memory of previous synthesized pixels (m_i), followed by a feed-forward sub-network. Positional encodings p_i are added in the first layer. (b) The operation performed in Local Self-Attention (example of a 2D case is shown). The image is partitioned into a grid of spatial blocks known as query blocks. In the self-attention operation, each pixel in a query block attends to all pixels in the memory block (shown in cyan rectangle). White grid locations show masked inputs that have zero-contribution towards the self-attention.

Transformers for Image and Scene Generation

- The task of generating realistic images from text.
- DALL·E takes as input a single stream of 1280 tokens (256 for the text and 1024 for the image), and is trained to generate all other tokens autoregressively (one after another)



(a) (b) (c) (d) (e) (f) (g)

Fig. 10: Images generated by DALL·E [20] from the following text prompts. (a) *An armchair in the shape of an avocado.* (b) *A photo of San Francisco's golden gate bridge.* Given a part of the image (in green box), DALL·E performs the image completion. (c) *An emoji of a baby penguin wearing a blue hat, red gloves, green shirt, and yellow pants.* (d) *An extreme close-up view of a capybara sitting in a field.* (e) *A cross-section view of a pomegranate.* (f) *A penguin made of watermelon.* (g) *The exact same cat on the top as a sketch on the bottom.*

Transformers for Low-level Vision

- Image super-resolution, Denoising, Deraining, and Colorization

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
Super-resolution	TTSR [16]	Texture enhancing Transformer module, Relevance embeddings to compute the relevance between the low-resolution and reference image.	2D Image	2D Image	Reconstruction loss, Perceptual loss defined on pretrained VGG19 features.

Texture Transformer for Super Resolution (TTSR)

- Transformer network for super-resolution

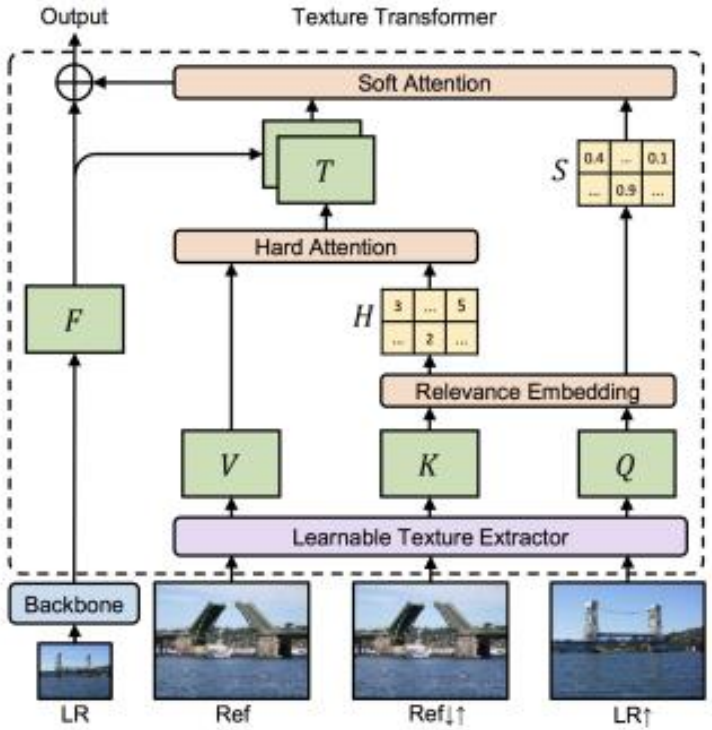


Fig. 11: Diagram of the texture Transformer module. Q (query), K (key) and V (value) represent texture features extracted from a (bicubic upsampled) low-resolution image, a sequentially down/upsampled reference image, and an original reference image, respectively. The relevance embedding aims to estimate similarity between low-resolution and reference images. H and S respectively denote hard and soft attentions computed from relevance embedding. T indicates high-resolution texture features that are then transferred to the features F of low-resolution image. Figure is from [16].

Transformers for Low-level Vision

- **Colorization Transformer**
- Self-attention used in the colorization Transformer is based on row/column attention layers introduced in these layers capture the interaction between each pixel of an input image while being computationally less costly.
- The row-wise attention layer applies self-attention to all pixels in a given row, while the column-wise attention layer considers pixels only in a given column of an image.

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
Image Colorization	ColTran [24]	Conditional Row/column multi-head attention layers, Progressive multi-scale colorization scheme.	2D Image	2D Image	Negative log-likelihood of the images

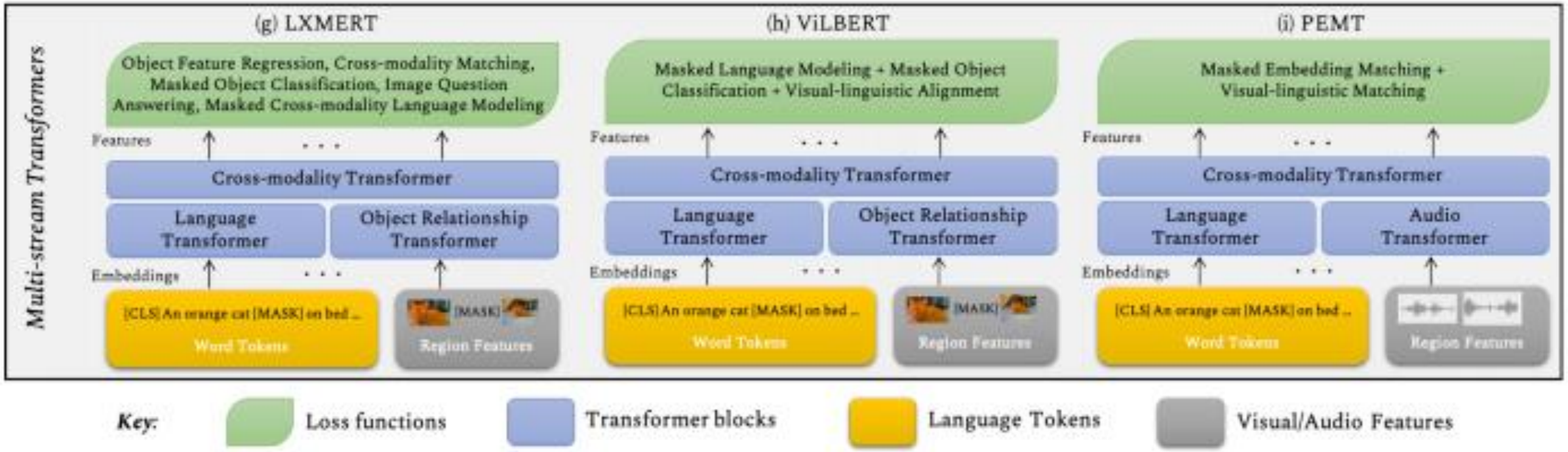
Figure 1: Samples of our model showing diverse, high-fidelity colorizations.

Transformers for Low-level Vision

Task	Method	Metric	Dataset	Performance	Highlights	Limitations
Super-Resolution	TTSR [16] CVPR'20	PSNR/ SSIM	CUFED5 Sun80 Urban100 Manga109	27.1 / 0.8 30.0 / 0.81 25.9 / 0.78 30.1 / 0.91	**a)** Achieves state-of-the-art super-resolution by using attention, **b)** Novel Transformer inspired architectures that can process multi-scale features.	**a)** Proposed Transformer does not process images directly but features extracted by a convolution based network, **b)** Model with large number of trainable parameters, and **c)** Compute intensive.
Image Coloriza-tion	ColTran [24] ICLR'21	FID	ImageNet	19.71	**a)** First successful application of Transformer to image colorization, **b)** Achieves SOTA FID score.	**a)** Lacks end-to-end training, **b)** limited to images of size 256×256.

Transformers for Multi-Modal Tasks

Multi-stream Transformers



The vokens (visualized tokens)

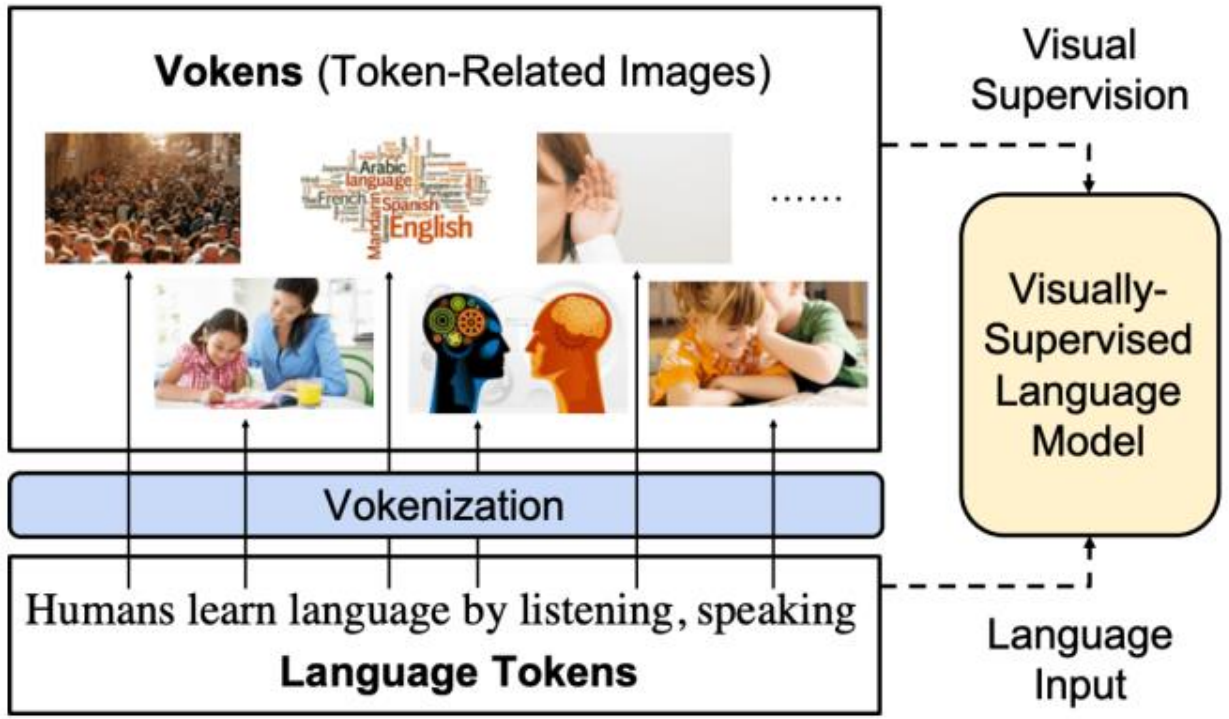
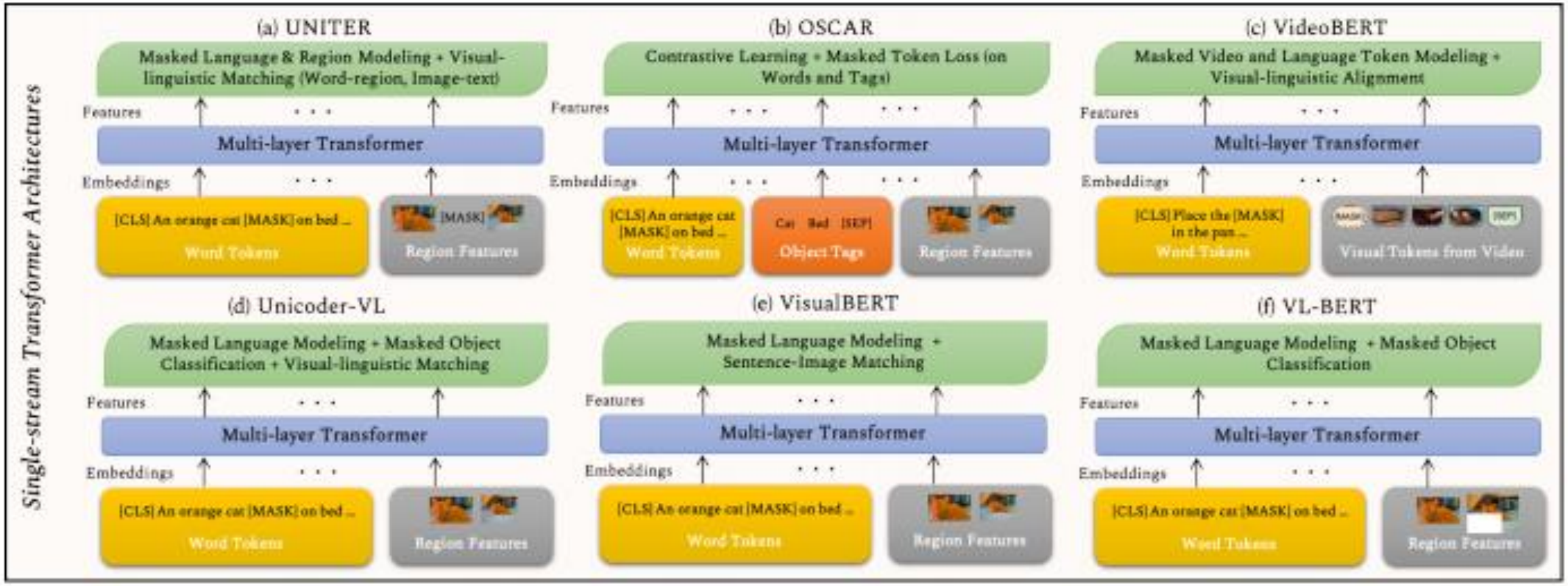


Fig. 13: Visualized tokens (Vokens) [191]: A language model is visually supervised using closely related images that leads to better feature representations from the pretrained model. Figure from [191].

Transformers for Multi-Modal Tasks

Single-stream Transformers



Transformers for Multi-Modal Tasks

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
Multi-Model Learning	Oscar [44]	Transformer layer to jointly process triplet representation of image-text [words, tags, features], Masked tokens to represent text data.	2D Image	Captions, Class labels, Object tags	Negative log-likelihood of masked tokens, Contrastive binary cross-entropy

Task	Method	Metric	Dataset	Performance	Highlights	Limitations
Multi-Model Learning	ViLBERT [181] NeurIPS'19	Acc./ mAP (R@1)	VQA [183]/ Retrieval [239]	70.6/ 58.2	**a)** Proposed Transformer architecture can combine text and visual information to understand inter-task dependencies, **b)** Achieves pre-training on unlabelled dataset.	**a)** Requires large amount of data for pre-training, **b)** Requires fine tuning to the new task.
	Oscar [44] ECCV'20	Acc./ mAP (R@1)	VQA [240]/ COCO	80.37/57.5	**a)** Exploit novel supervisory signal via object tags to achieve text and image alignment, **b)** Achieves state-of-the-art results.	Requires extra supervision through pre-trained object detectors thus performance is dependent on the quality of object detectors.
	UNITER [43] ECCV'20	Acc./ Avg. (R@1/5/10)	VQA [183]/ Flickr30K [241]	72.47/83.72	Learns fine-grained relation alignment between text and images	Requires large multi-task datasets for Transformer training which lead to high computational cost.

Transformers for 3D Analysis

- The Mesh Transformer (METRO)

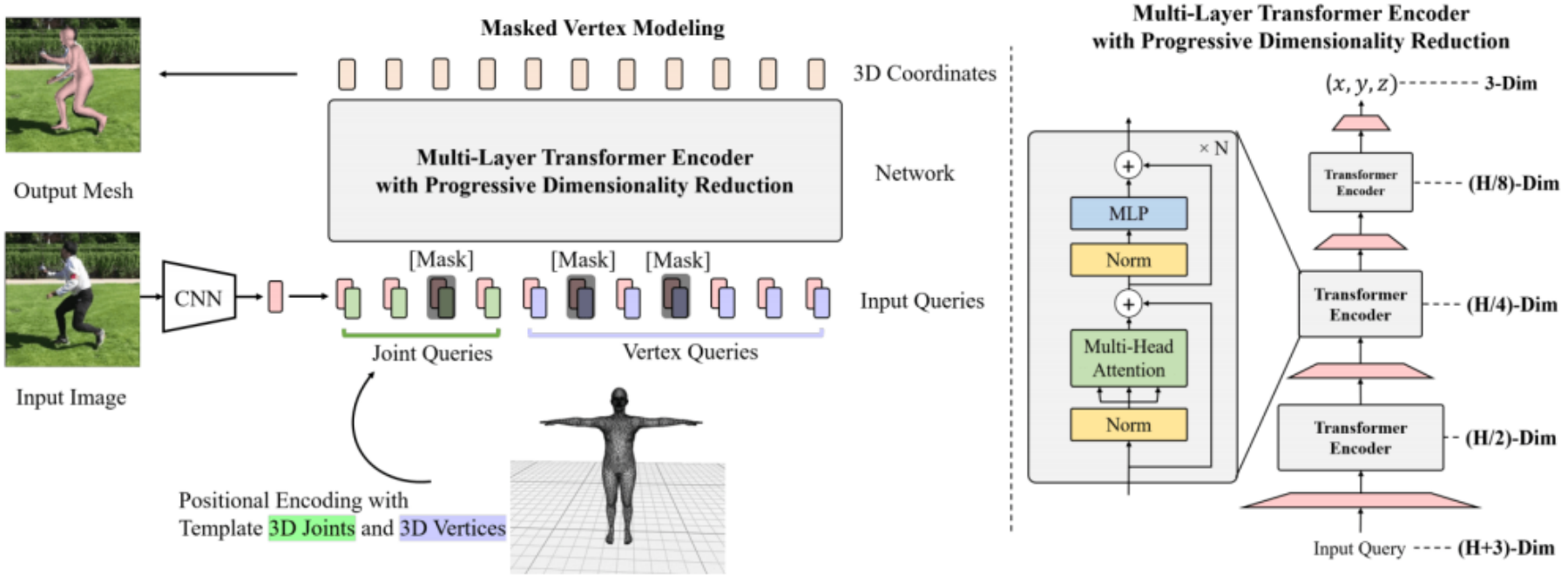


Fig. 16: Mesh Transformer architecture. The joint and vertex queries are appended with positional embeddings and passed through multiple self-attention layers to jointly regress 3D coordinates of joints and mesh vertices. Figure is from [45].

Transformers for 3D Analysis

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
3D Classification/Segmentation	PT [230]	Point Transformer block, Transition down block to reduce cardinality of the point set, Transition up for dense prediction tasks.	CAD models, 3D object part segmentation	Object and shape categories	Cross-entropy
3D Mesh Reconstruction	METRO [45]	Progressive dimensionality reduction across Transformer layers, Positional Encoding with 3D joint and 3D vertex coordinates, Masked vertex/joint modeling.	2D Image	3D Mesh + Human Pose	L_1 loss on mesh vertices and joints in 3D and 2D projection.

Task	Method	Metric	Dataset	Performance	Highlights	Limitations
3D Analysis	Point Transformer [230] arXiv'20	Top-1 Acc. IoU	ModelNet40 [232]	92.8 85.9	**a)** Transformer based attention capable to process unordered and unstructured point sets, **b)** Permutation invariant architecture.	**a)** Only moderate improvements over previous SOTA, **b)** Large number of trainable parameters around 6× higher than PointNet++ [242].
	METRO [45] arXiv'20	MPJPE PA-MPJPE MPVE	3DPW [235]	77.1 47.9 88.2	**a)** Does not depend on parametric mesh models so easily extendable to different objects, **b)** Achieves SOTA results using Transformers.	Dependent on hand-crafted network design.

OPEN CHALLENGES & FUTURE DIRECTIONS

- High Computational Cost
- Large Data Requirements
- Vision Tailored Transformer Designs
- Neural Architecture Search for ViTs
- Interpretability of Transformers
- Hardware Efficient Designs
- Towards Integrating All Modalities

Q&A