Computational Visual Media
https://doi.org/10.1007/s41095-021-0247-3 Vol. 8, No. 1, March 2022, 33–62

Review Article

Transformers in computational visual media: A survey

Yifan Xu[1,2], Huapeng Wei[3], Minxuan Lin[1,2], Yingying Deng[1,2], Kekai Sheng[4], Mengdan Zhang[4], Fan Tang[3], Weiming Dong[1,2,5] (✉), Feiyue Huang[4], and Changsheng Xu[1,2,5]

© The Author(s) 2021.

Abstract Transformers, the dominant architecture for natural language processing, have also recently attracted much attention from computational visual media researchers due to their capacity for long-range representation and high performance. Transformers are sequence-to-sequence models, which use a self-attention mechanism rather than the RNN sequential structure. Thus, such models can be trained in parallel and can represent global information. This study comprehensively surveys recent visual transformer works. We categorize them according to task scenario: backbone design, high-level vision, low-level vision and generation, and multimodal learning. Their key ideas are also analyzed. Differing from previous surveys, we mainly focus on visual transformer methods in low-level vision and generation. The latest works on backbone design are also reviewed in detail. For ease of understanding, we precisely describe the main contributions of the latest works in the form of tables. As well as giving quantitative comparisons, we also present image results for low-level vision and generation tasks. Computational costs and source code links for various important works are also given in this survey to assist further development.

1 NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: Y. Xu, xuyifan2019@ia.ac.cn; M. Lin, linminxuan2018@ia.ac.cn; Y. Deng, dengyingying2017@ia.ac.cn; W. Dong, weiming.dong@ia.ac.cn (✉); C. Xu, changsheng.xu@ia.ac.cn.

2 School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100040, China.

3 School of Artificial Intelligence, Jilin University, Changchun 130012, China. E-mail: H. Wei, weihp20@jlu.edu.cn; F. Tang, tangfan@jlu.edu.cn.

4 Youtu Lab, Tencent Inc., Shanghai 200233, China. E-mail: K. Sheng, saulsheng@tencent.com; M. Zhang, davinazhang@tencent.com; F. Huang, garyhuang@tencent.com.

5 CASIA-LLVISION Joint Lab, Beijing 100190, China.

Manuscript received: 2021-06-17; accepted: 2021-07-16

Keywords visual transformer; computational visual media (CVM); high-level vision; low-level vision; image generation; multi-modal learning

1 Introduction

Convolutional neural networks (CNNs) [1–3] have become the fundamental architecture in computational visual media (CVM). Researchers began to incorporate a self-attention mechanism into CNNs to model long-range relationships, due to the problem of locality of convolutional kernels [4–8]. Recently, Dosovitskiy et al. [9] found that using a self-attention-only structure, without convolution, works well in computer vision. Since then, the transformer architecture [10], a non-convolutional architecture dominating the research field of natural language processing (NLP), has has been used in computer vision. Introducing transformers into computer vision provides four advantages that CNNs lack:

- Transformers learn with more inductive bias and performs better when trained on large datasets (e.g., ImageNet-21K or JFT-300M) [9, 11].
- Transformers provide a more general architecture suitable for most fields, including NLP, CV, and multimodal learning.
- Transformers powerfully model long-range interactions in a computationally-efficient manner [12, 13].
- The learned representation of relationships is more general and robust than the local patterns from convolution modules [14].

As Table 1 shows, an increasing number of works on visual transformers have come out in various subfields of computational visual media. An instructive survey is important because of the difficulties in arranging

Table 1 Recent visual transformers introduced in this survey

Area	Secondary area	Method	Contributions
Backbone network	Classification	T2T ViT [15]	An effective and efficient tokens-to-token module
		TNT [16]	The first to exploit the benefit of pixel-level relations
		CPVT [17]	An instance-level position embedding module
		ConViT [18]	Adaptive reception field in visual transformers
		DeepViT [19]	A Re-Attention module for deep-layer ViTs
		Swin Transformer [20]	A shifted-window based MSA & a deep-narrow module
		PiT [21]	The first to investigate the benefit of pooling in ViTs
		LocalViT [22]	A depth-wise convolution based module to exploit locality
	Visualization	Transformer-Explainability [23]	A better tool to visualize feature maps from ViT models
High-level vision	Detection	DETR [24]	First transformer-based detection SOTA model
		Deformable DETR [25]	An efficient attention module reducing time consumption
		UP-DETR [26]	An unsupervised pre-training method for DETR
		PVT [27]	A general transformer architecture for dense prediction
	Segmentation	VisTR [28]	First transformer-based segmentation model
		SegFormer [29]	A lightweight efficient segmentation transformer model
Low-level vision	Colorization	ColTran [30]	First transformer-based image colorization model
	Text-to-image	TIME [31]	Text-to-image generation
		DALL·E [32]	Zero-shot text-to-image generation framework
	Super resolution	IPT [11]	Image processing model
		TTSR [33]	Flexible application of transformer
	Image generation	TransGAN [34]	First pure transformer-based GAN for generation
		GANsformer [35]	A bipartite transformer
		VQGAN [36]	A transformer-based high-resolution image generator
	Image restoration	Uformer [37]	A transformer-based hierarchical encoder–decoder network
	Style transfer	StyTr^2 [38]	First transformer-based style transfer model
	Point cloud learning	PCT [39]	Among the first transformer-based point cloud models
Multi-modality learning	Two-stream model	ViLBERT [40]	The first proposed two-stream model for V+L tasks
	Single-stream model	UNITER [41]	A universal model for joint multi-modal embedding
	Mixed model	SemVLP [42]	First mixed single- and two-stream model

such fast and abundant developments. Due to the fast development of visual transformer backbones, this survey specifically focuses on the latest works in that area, as well as low-level vision tasks.

Specifically, this study is mainly arranged into four specific fields: backbone design, high-level vision (e.g., object detection and semantic segmentation), low-level vision and generation, and multimodal learning. We highlight backbone design and low-level vision as our main focus in Fig. 1. The developments to be introduced are summarised in Table 1. For backbone design, several latest works are introduced, considering two aspects: (i) injecting convolutional prior knowledge into ViT, and (ii) boosting the richness of visual features. We also summarize the breakthrough ideas of each work in Fig. 1. For high-level vision, we introduce the mainstream of DETR-based transformer detection models [24]. For low-level vision and generation, we arrange papers according to different subareas including colorization [30, 43–45], text-to-image [31, 32, 46], super-resolution [47–49], and image generation [50–54]. For multimodal learning, we review some recent representative works on vision-plus-language (V+L) models and summarize pretraining objectives in this field.

We comprehensively compare results in different fields and give training details, including computational cost and source code links to facilitate and encourage further research. Some images resulting from low-level vision models are also illustrated. The rest of the paper is organized as follows. Section 2 introduces visual transformers. Section 3 lists latest

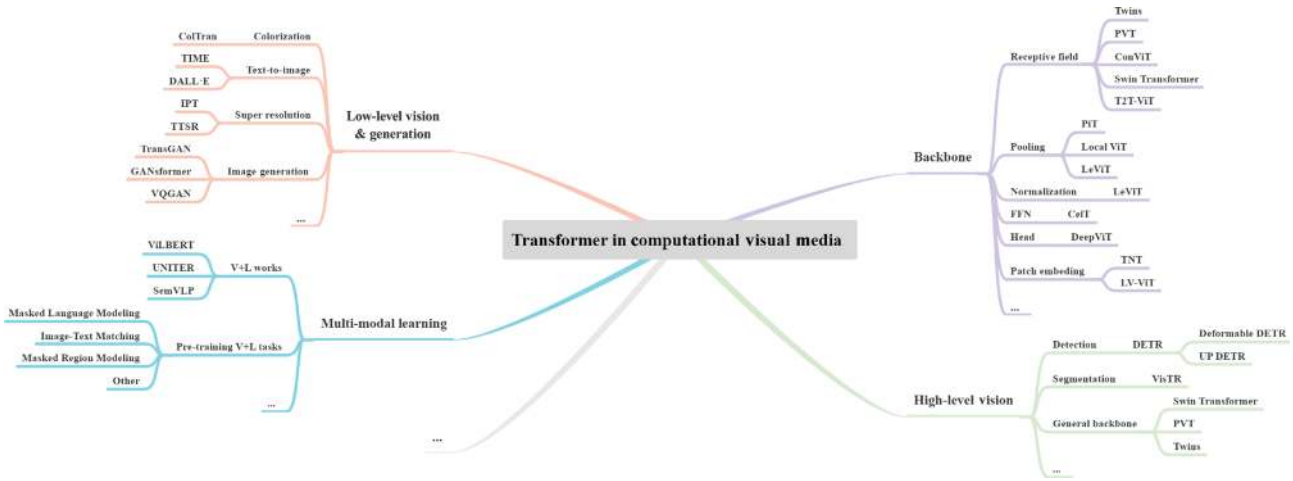

Fig. 1 Organisation of recent works on visual transformers.

developments in backbone networks for visual transformers in image classification. Section 4 describes several recent advanced designs using visual transformers in object detection. Section 5 introduces transformer-based methods for various low-level vision tasks. Section 6 reviews recent representative works on multimodal learning. Finally, we draw conclusions from different research fields in Section 7.

2 Visual transformers

Before introducing the latest developments, we give the basic formulation of visual transformers by using ViT [9] as an example. As shown in Fig. 2, a typical ViT mainly contains five basic procedures: splitting input images into smaller local patches, preparing the input token (patch tokens, class token, and position embedding), a series of stacked transformer blocks [55] (i.e., layer normalization (LN) [56] + multi-head self-attention (MSA) [57] + skip-connection layer [1] + multilayer perception (MLP) or feed-forward network (FFN)), and post-process module.

Formally, given an input image $\boldsymbol{X} \in \mathbb{R}^{H\times W\times C}$ and its labels $\boldsymbol{Y}$, $\boldsymbol{X}$ is first reshaped into a sequence of flattened 2D image patches $\boldsymbol{X}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$. Then, following BERT [10], a `class` token and several position tokens are used to record extra meaningful information for inference. Together, the input is formulated as follows:

$$z_0 = [\boldsymbol{x}_{\text{cls}}; \boldsymbol{x}_p^1\cdot\boldsymbol{E}; \cdots; \boldsymbol{x}_p^N\cdot\boldsymbol{E}] + [\boldsymbol{E}_{\text{pos}}^{\text{cls}}; \boldsymbol{E}_{\text{pos}}^1; \cdots; \boldsymbol{E}_{\text{pos}}^N]$$

where $\boldsymbol{x}_{\text{cls}} \in \mathbb{R}^D$ is the class token, $\boldsymbol{E} \in \mathbb{R}^{(P^2\cdot C)\times D}$ is a linear projection of each patch $\boldsymbol{X}_p$, and $\boldsymbol{E}_{\text{pos}}^i \in \mathbb{R}^D$ is the learnable position embedding for the i-th token.

Then, the input is sent into several sequential

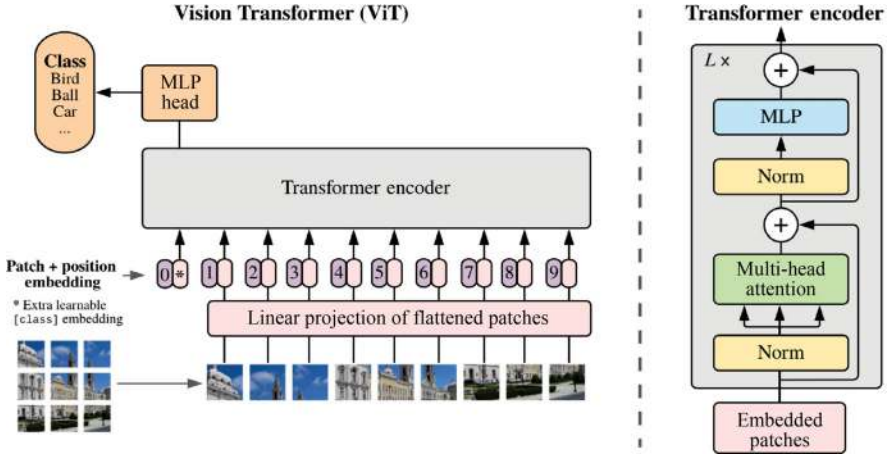

Fig. 2 Framework of ViT (left) and typical pipeline of a transformer encoder (right). Reproduced with permission from Ref. [9], © The Author(s) 2021.

transformer blocks:

$$z'_{l+1} = z_l + \mathrm{MSA}(\mathrm{LN}(z_l))$$
$$z_{l+1} = z'_{l+1} + \mathrm{MLP}(\mathrm{LN}(z'_{l+1}))$$

where $l \in \{0, \cdots, L-1\}$ denotes the layer, L is the number of transformer blocks, the MLP includes two fully-connected layers using GELU [58] as the activation function, $\mathrm{LN}(\cdot)$ is a layer-normalization module [56], and the MSA module is formulated as

$$\mathrm{MSA}(z) = [\mathrm{SA}_1(z); \cdots; \mathrm{SA}_H(z)] \times \boldsymbol{U}_{\mathrm{msa}}$$
$$\mathrm{SA}_i(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sigma\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \cdot \boldsymbol{V}$$

where z is the input, $[\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}] = z \times \boldsymbol{U}^i_{qkv}$, $\boldsymbol{U}^i_{qkv} \in \mathbb{R}^{D \times (3 \cdot D_h)}$ projects the D-dimensional input z to D_h-dimensional $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ in the head i, $\sigma(\cdot)$ is the softmax function, and $\boldsymbol{U}_{\mathrm{msa}} \in \mathbb{R}^{(H \cdot D_h) \times D}$ re-casts the output from H heads of the MSA module into one D-dimensional output. Several variants of MSA, like Reformer [59], Performer [60], and LinFormer [61], are available.

3 Backbone design

In this section, we describe several recent designs for the backbone of ViT models. Without loss of generality, we focus on the image classification task. We divide recent progress into two mainstream approaches: (i) injecting convolutional prior knowledge into ViT, works including T2T-ViT [15], ConViT [18], PiT [21], and Swin Transformer [20], and (ii) boosting the richness of visual features, including TNT [16], CPVT [17], DeepViT [19], and LocalViT [22]. We also briefly describe recent developments in visualizing feature maps of ViT models [23, 62, 63], which help to better understand the working mechanism of ViT models. We list core details of their performance on ImageNet [64] in Table 2.

3.1 Latest developments

3.1.1 T2T-ViT

Yuan et al. [15] note that the method to convert input images into tokens in a typical ViT [9] ineffectively models the spatial structure of image data and may lead to poor training efficiency and suboptimal performance. They propose two effective approaches to address the aforementioned problem. First, they propose a token-to-token (T2T) module to inject spatial information into the tokenization of image patches and reduce the length of tokens progressively for the sake of computational and parameter efficiency. Inspired by CNN architectures [1–3], they also devise a deep-narrow ViT framework to reduce the number of parameters and enhance training efficiency. Overall, they train ViT models from scratch on ImageNet without additional datasets.

3.1.2 TNT

Han et al. [16] propose a novel Transformer-iN-Transformer (TNT) framework to further exploit the intrinsic spatial structural information in image data. As Fig. 3 shows, TNT considers patch and pixel level relations in learning useful visual features. They propose a TNT block to utilize the pixel-level representations effectively and efficiently. They introduce an additional transformer called an Inner T-Block to model pixel-level relationships in each patch and then reinforce the patch-level features with the calculated pixel-level ones. Consequently, TNT achieves 81.3% top-1 classification accuracy on ImageNet [64] at the cost of only moderate additional computation. The experimental results verify the positive effects of pixel-level relation modeling.

3.1.3 ConViT

D'Ascoli et al. [18] propose a novel ViT model with soft convolutional inductive biases (ConViT) to endow transformers with an adaptive receptive field. Figure 4 schematically shows the core block, called a gated positional self-attention (GPSA) module. A GPSA block has two branches: $\boldsymbol{W}_{\mathrm{qry}}$ or $\boldsymbol{W}_{\mathrm{key}}$ is used to model the global or long-range relationship, and $\boldsymbol{v}_{\mathrm{pos}}$ is utilized to model the relationship within local regions. To adaptively trade-off between the two branches, they adopt a learnable parameter λ, which

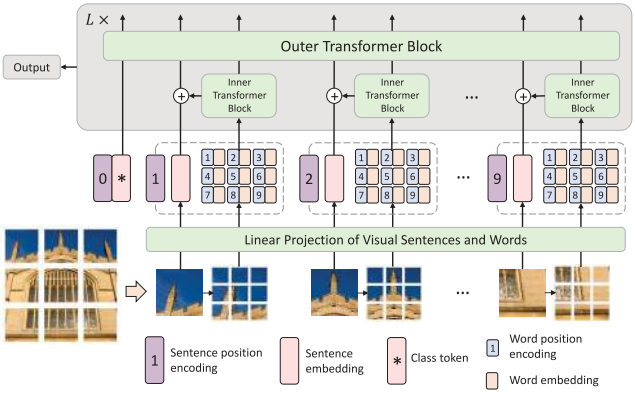

Fig. 3 Framework of TNT. Reproduced with permission from Ref. [16], © The Author(s) 2021.

Table 2 Classification accuracy on ImageNet [64] for various visual transformers

Method	Image size	FLOGs (G)	#Param (M)	Acc (%)	Source (GitHub)
			Convolution-based neural network		
ResNet [1]	224^2	4.1	25.6	76.2	—
RegNetY-4G [3]	224^2	4.0	21	80.0	`facebookresearch/pycls`
RegNetY-16G [3]	224^2	16.0	84	82.9	
EfficientNet-B0 [2]	224^2	0.4	5.3	77.1	`rwightman/gen-efficientnet-pytorch`
EfficientNet-B1 [2]	224^2	0.7	7.8	79.1	
EfficientNet-B3 [2]	300^2	1.8	12	81.6	
EfficientNet-B5 [2]	456^2	9.9	30	83.6	
EfficientNet-B7 [2]	600^2	37.0	66	84.3	
			Visual transformer		
ViT [9]	384^2	55.4	86	77.9	`google-research/vision_transformer`
	384^2	190.7	307	76.5	
DeiT [65]	224^2	4.6	22	79.8	`facebookresearch/deit`
	384^2	55.4	86	83.1	
T2T ViT [15]	224^2	5.2	21.5	80.7	`yitu-opensource/T2T-ViT`
TNT [16]	224^2	5.2	23.8	81.3	`huawei-noah/noah-research/tree/master/TNT`
	224^2	14.1	65.6	82.8	
CPVT [17]	224^2	—	23	81.5	`Meituan-AutoML/CPVT`
	224^2	—	88	82.3	
ConViT [18]	224^2	5.4	27	81.3	—
	224^2	17	86	82.4	
DeepViT [19]	224^2	—	27	82.3	`zhoudaquan/dvit_repo`
	224^2	—	55	83.1	
Swin Transformer [20]	224^2	4.5	29	81.3	`microsoft/Swin-Transformer`
	384^2	47.0	88	84.2	
PiT [21]	224^2	4.6	22.1	81.9	`naver-ai/pit`
	224^2	12.5	73.8	84.0	
LocalViT [22]	224^2	4.6	22.4	80.8	`ofsoundof/LocalViT`

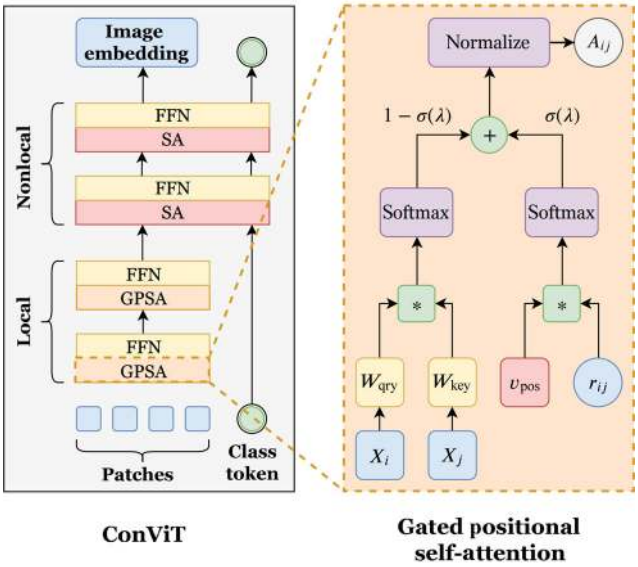

Fig. 4 Framework of ConViT and the gated positional self-attention mechanism. Reproduced with permission from Ref. [18], © The Author(s) 2021.

is initialized as 1 for all layers and all heads in MSA. With the proposed GPSA module, they manage to adaptively expand the self-attention receptive field during training.

3.1.4 CPVT

Chu et al. [17] resort to a novel design of position embedding module to further reinforce the richness of learned visual features from ViT. Instead of a predefined position embedding that is independent of the input data, they propose a conditional position embedding scheme to generate different positional encodings for various input tokens, akin to dynamic neural network design [66]. In their implementation, they also rearrange the input tokens in a spatial manner and apply convolution operations to extract the position embedding in a learnable way. In this way, they also maintain the local neighborhood information during tokenization, benefiting classification performance.

Two further ViT models, LeViT [12] ① and CoaT [67] ②, investigate the importance of position embedding and propose different implementations. We do not describe them further due to lack of space.

3.1.5 Swin Transformer

On the basis of the observations that image data contain much redundant spatial information and given the success of deep-narrow CNN architectures, Liu et al. [20] propose a novel hierarchical visual transformer design. Figure 5(a) illustrates the core idea of the window MSA (W-MSA) and the shifted W-MSA (SW-MSA) within Swin Transformer, which separate local patches into several windows and run the MSA module window by window. With the W-MSA mechanism, they reduce the computation complexity from $\mathcal{O}(4HWC^2+2(HW)^2C)$ to $\mathcal{O}(4HWC^2 + 2M^2HWC)$, where H and W represent the size of input patches, $M \times M$ is the number of windows, and C is the feature dimension. A shifted window design is also proposed to encourage cross-window communication for rich visual features. They also propose a deep-narrow architecture (see Fig. 5(b)). Extensive experiments on ImageNet, COCO, and ADE-20K demonstrate that Swin Transformer enhances efficient use of parameters and achieves state-of-the-art object detection and semantic segmentation.

Dong et al. [68] also propose another vision transformer model, CSWin Transformer, which utilizes a cross-shaped window self-attention mechanism (akin to criss-cross attention [69] or strip pooling [70]) and a locally enhanced position encoding. CSWin Transformer obtains even better performance than SWin Transformer.

3.1.6 DeepViT

Layer scaling (e.g., 152-layer ResNet [1]) is an important aspect of CNN architectures. With regard to ViT models, Zhou et al. [19] empirically find that the performance of deep layer ViT models saturates when we stack more than 20 transformer blocks even with the help of skip-connection layers. They unveil that the reason is attention collapse: the feature maps extracted from each head in one MSA module share increasingly similar patterns, leading to huge information redundancy and low training efficiency. If the communication between the MSA heads is promoted, the information redundancy between each head and rich learned visual feature can be reduced.

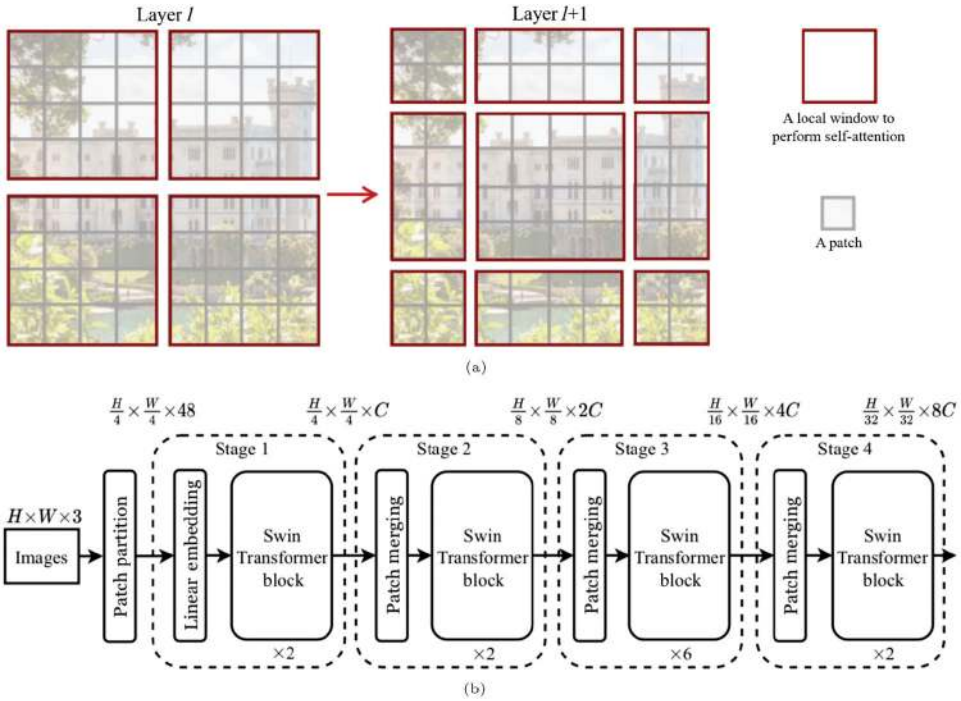

Fig. 5 (a) Window MSA (W-MSA) greatly reduces computational cost and facilitates communication between each isolated W-MSA. (b) Overview of Swin Transformer. Reproduced with permission from Ref. [20], © The Author(s) 2021.

① https://github.com/facebookresearch/LeViT

② https://github.com/mlpc-ucsd/CoaT

On the basis of the aforementioned motivation, they propose a simple and effective Re-Attention module:

$$Norm\left(\Theta^{\mathrm{T}} softmax\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\right)\cdot\boldsymbol{V}$$

where $\Theta \in \mathbb{R}^{H\times H}$ is a learnable parameter to facilitate the communication between the H heads within one MSA module. Experiments on ImageNet [64] verify that a 32-layer ViT model can be trained without performance saturation with the help of a Re-Attention module.

Notably, concurrent work, which is termed CaiT [71], also investigates the topic of layer scaling and proposes a different perspective. Further details can be obtained from their paper.

3.1.7 PiT

Considering the importance of the pooling layer to model capability and generalization performance of CNN architectures, Heo et al. [21] investigate the possibility of taking advantage of pooling modules in ViT. The pooling layer in a conventional CCN architecture conducts spatial information aggregation for spatially invariant features. On the basis of this observation, they propose to implement spatial information condensation via depth-wise convolution. As shown in Fig. 6, they first split the obtained input tokens into class tokens and spatial ones, and then they recover the spatial shape of the latter. Next, they leverage a depth-wise convolution operation on the spatial branch for the purpose of a pooling layer. Meanwhile, they apply a fully connected layer to project the class token into the same dimension. With a simple and effective pooling module, they propose a pooling-based ViT (PiT) and achieve an optimal trade-off between computation efficiency and classification performance.

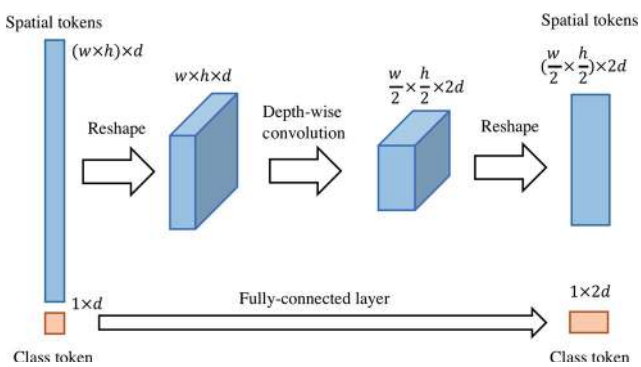

Fig. 6 Pooling layer in the PiT architecture. Reproduced with permission from Ref. [21], © The Author(s) 2021.

3.1.8 LocalViT

Li et al. [22] study the differences between ViT models and CNN architectures. They find that visual transformers are good at modeling global relations while lacking a local scheme to learn interactions within a local region, which is the characteristic of convolution. A local mechanism is important and useful for modeling spatial structures for image data. Thus, they believe that visual transformers must reinforce the model's capability for local relation modeling to promote the learned visual features from ViT models. Specifically, they investigate several possible blocks and then propose local ViT (LocalViT), as shown in Fig. 7(right). Experiments on ImageNet [64] indicate that the LocalViT module is a practical local mechanism which boosts the performance of various ViT models [15, 16, 27, 65].

3.2 Comparison on ImageNet

We compare the classification accuracies of the latest ViT models on the ImageNet benchmark [64] in Table 2, together with their implementation details, namely, #FLOGs, #Param, and source code, to facilitate further research. The experimental values indicate that ViT models have potential to achieve comparable performance or even outperform state-of-the-art CNN architectures like RegNet [3] and EfficientNet [2], which are based on expert-designed basic modules and the power of neural architecture search (NAS) techniques. We also observe very recent exciting progress, in that the latest proposed ViT models possess higher model capability and better parameter efficiency for vision than the original version of ViT.

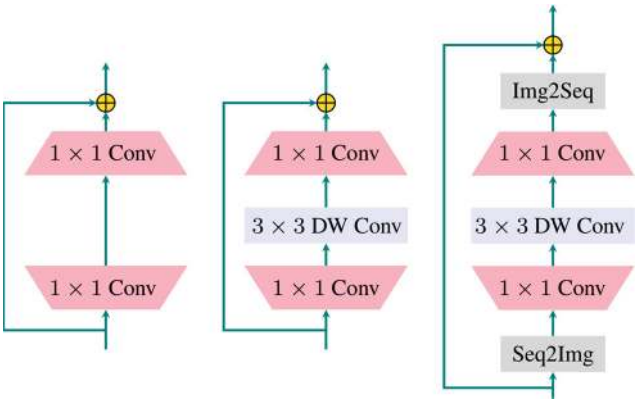

Fig. 7 Comparison of the convolutional version of the FFN module in ViT models (left), inverted residual blocks (center), and the proposed module in LocalViT to exploit the benefit of locality in ViT (right). Reproduced with permission from Ref. [22], © The Author(s) 2021.

3.3 Visualization of ViT

Visualizing the feature maps in ViT is also an interesting and worthy research topic. As ViT models leverage different basic components from CNN models, we should adopt different visualization methods correspondingly. As shown in Fig. 8, the latest tools specialized for MSA modules and ViT models, namely, partial LRP [63] and Transformer-Explainability [23][①], can generate better results for feature map visualization than the visualization methods for CNN. The visualizations indicate that ViT models can learn additional meaningful spatial information with image-level annotations alone. Therefore, ViT models have potential values in weakly supervision scenarios, such as weakly supervised object detection.

4 High-level vision

In this section, we focus on representative recent high-level vision tasks based on transformer framework. High-level vision refers to stages of visual processing that transition from analyzing local image structure to exploring the structure of the external world that produced those images. The main tasks include object detection [24–26], segmentation [28, 29, 74–79], and key-point detection [80–85]. As the focus of this survey is low-level vision tasks, we only briefly introduce some interesting works in object detection. Modern detection methods address the set prediction task by defining a large set of proposals [86, 87], anchors [88], or window centers [89, 90]. Unlike previous attempts [91–96], transformer-based detection raises the possibility of total anchor-free and end-to-end models. We begin with the stream of DETR [24], followed by Deformable DETR [25] and UP-DETR [26]. A more complete approach, PVT [27], which is the earliest transformer backbone for dense prediction tasks like detection, is also introduced. Additional recent high-level backbones like Swin Transformer [20] and Twins [97] are introduced in Section 3. A comparison is provided in Table 3.

4.1 DETR

Carion et al. [24] were the first to provide a completely end-to-end detection model based on the transformer encoder–decoder architecture. It gives researchers a new insight that the transformer architecture can achieve state-of-the-art performance in detection. Unlike previous detection models, DETR does not rely on artificially designed anchors. The overall structure is illustrated in Fig. 9. The transformer encoders are arranged after a convolution feature

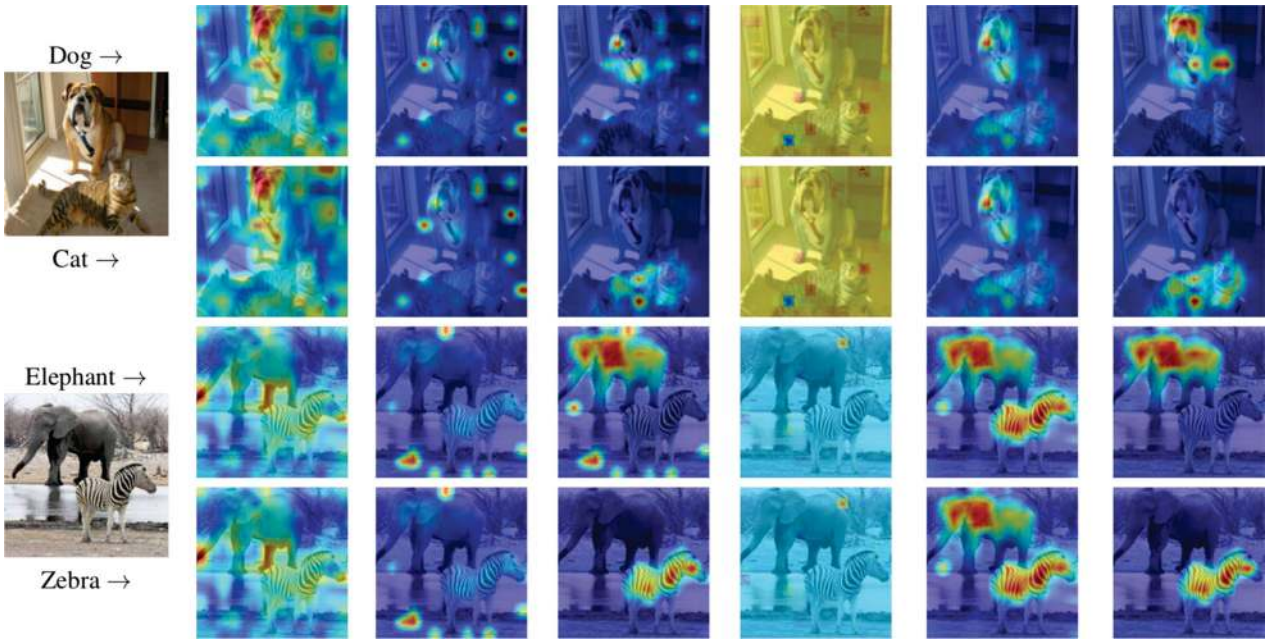

Fig. 8 Class-specific visualization results from ViT. Left to right: input image, rollout [62], raw-attention, GradCAM [72], LRP [73], partial LRP [63], and Transformer-Explainability [23]. Reproduced with permission from Ref. [23], © The Author(s) 2021.

① https://github.com/hila-chefer/Transformer-Explainability

Table 3 Comparison of transformer-based detection models on the COCO 2017 val set

Method	Training epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	#Param (M)	FPS	FLOPS (G)	Source (GitHub)
Convolution-based models											
FCOS [90]	36	41.0	59.8	44.1	26.2	44.6	52.2	—	23	177	`tianzhi0549/FCO`
Faster R-CNN+FPN [86]	109	42.0	62.1	45.5	26.6	45.4	53.4	42	26	180	`rbgirshick/py-faster-rcnn`
Transformer-based models											
DETR [24]	500	42.0	62.4	44.2	20.5	45.8	61.1	41	28	86	`facebookresearch/detr`
DETR-DC5 [24]	500	43.3	63.1	45.9	22.5	47.3	61.1	41	12	187	
Deformable DETR [25]	50	43.8	62.6	47.7	26.4	47.1	58.0	40	19	173	`fundamentalvision/Deformable-DETR`
UP-DETR [26]	150	40.5	60.8	42.6	19.0	44.1	60.0	41	—	—	`dddzg/up-detr`
UP-DETR [26]	300	42.8	63.0	45.3	20.8	47.1	61.7	41	—	—	
PVT-T [27]	300	36.7	56.9	38.9	22.6	38.8	50.0	23.0	—	—	`whai362/PVT`
PVT-M [27]	300	41.9	63.1	44.3	25.0	44.9	57.6	53.9	—	—	
ViT-B/16-FRCNN [98]	21	37.8	57.4	40.1	17.8	41.4	57.3	21	—	—	—

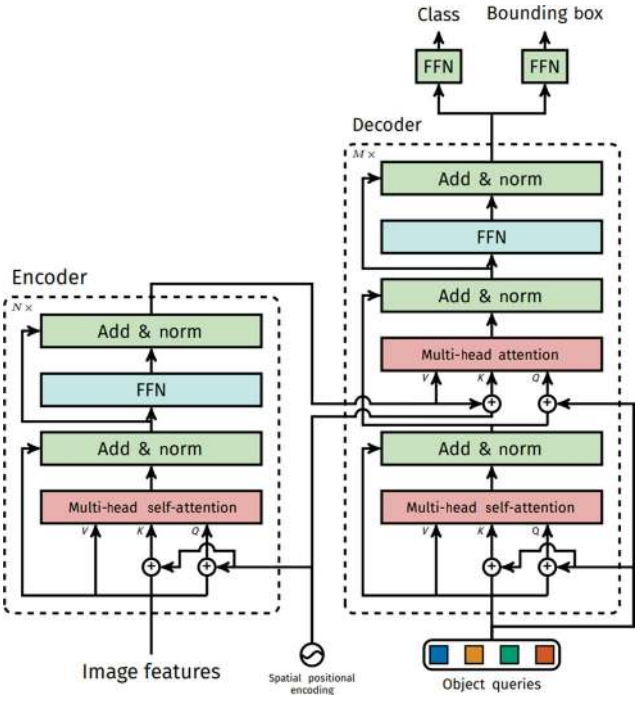

Fig. 9 Overall structure of DETR. Reproduced with permission from Ref. [24], © Springer Nature Switzerland AG 2020.

extractor. We introduce the encoder and decoder modules in order.

In the DETR encoder, first, the output feature map of CNN is decomposed into patches as for ViT [9] introduced in Section 3. Then, the patches are mapped to one-dimensional vectors to go through several traditional transformer encoders. DETR and traditional BERT encoders differ only in the positional embedding. The positional embedding is injected into all encoder blocks rather than only the input layer to preserve the positional information; they claim that high-level vision detection needs more positional information than classification. Only queries and keys are also injected into the positional embedding.

The decoder has two inputs. The first is the object query, which only serves as queries in the second MSA layers. The second is the output of the encoder module, which serves as values and keys for the second MSA layers. To easily understand the mechanism, readers can treat the object queries as information of different target objects and suppose the decoder aims to find whether the similar patterns to the object queries exist in the image features. The output of the decoder is then passed through two branches, namely, the box and class branches. The box branch predicts the positions of the target objects, while the class branch serves to predict the category of each predicted box.

4.2 Deformable DETR

Although considerable progress has been achieved by DETR in transformer-based detection, its main deficiency is its huge computational cost. Training a DETR on one V100 GPU is reported to take 48 days, which is unaffordable for common institutions. Thus, Zhu et al. [25] propose the deformable self-attention module to reduce the training time to 340 GPU hours at the same time as improving the original performance. The core idea of deformable attention is to find the nearest K values of an input query to

calculate attention. Nearest here refers to semantic distance rather than spatial distance. Deformable attention is illustrated in Fig. 10, which is drawn from deformable convolution [99]. A linear model is established to learn the offsets of the nearest K values, and then another linear model is established to learn the attention score of each value. In summary, the main contributions of deformable attention module are (i) only K corresponding values rather than all values are required to calculate the attention of one query and (ii) the attention scores are learned by a network rather than by simple multiplication of queries and keys.

4.3 UP-DETR

Dai et al. [26] propose a novel unsupervised pretraining method called random query patch detection for DETR [24, 25], which leads to better performance. Figure 11(a) illustrates their pretraining method. A random query patch is randomly cropped from an input image. Then, the query patch is added to the object queries of the DETR decoder. The final goal is to predict two things: (i) $\mathcal{L}_{\text{cls}}$, that is, existence of objects in the query patch, and (ii) $\mathcal{L}_{\text{box}}$, that is, the location of the query patch in the image. A reconstruction loss $\mathcal{L}_{\text{rec}}$ is also designed to ensure that the CNN has extracted full information from the query patch. This pretraining method leads to more flexible training. As shown in Fig. 11(b), a more robust representation can be learned after augmenting the random query patches.

4.4 PVT

Diverging from DETR, Wang et al. [27] also propose a pure transformer-based backbone, called a pyramid vision transformer (PVT), for detection and segmentation. Its framework is shown in Fig. 12. After each stage, the output is rearranged to recover spatial structure and is then down-sampled to half resolution. Notably, the spatial reduction is only conducted on the key $\boldsymbol{K}$ and value $\boldsymbol{V}$ while the spatial size of the query $\boldsymbol{Q}$ is maintained. In practice, the full architecture of PVT-based detection models includes a PVT backbone and a general detection head, such as RetinaNet [88] and Mask R-CNN [100]. Recently, several dense prediction backbones have come out after PVT [27], like Swin Transformer [20], CPVT [17], and Twins [97], which we introduced in Section 3.

5 Low-level vision and generation

In this section, we focus on some representative recent transformer-based works on low-level vision

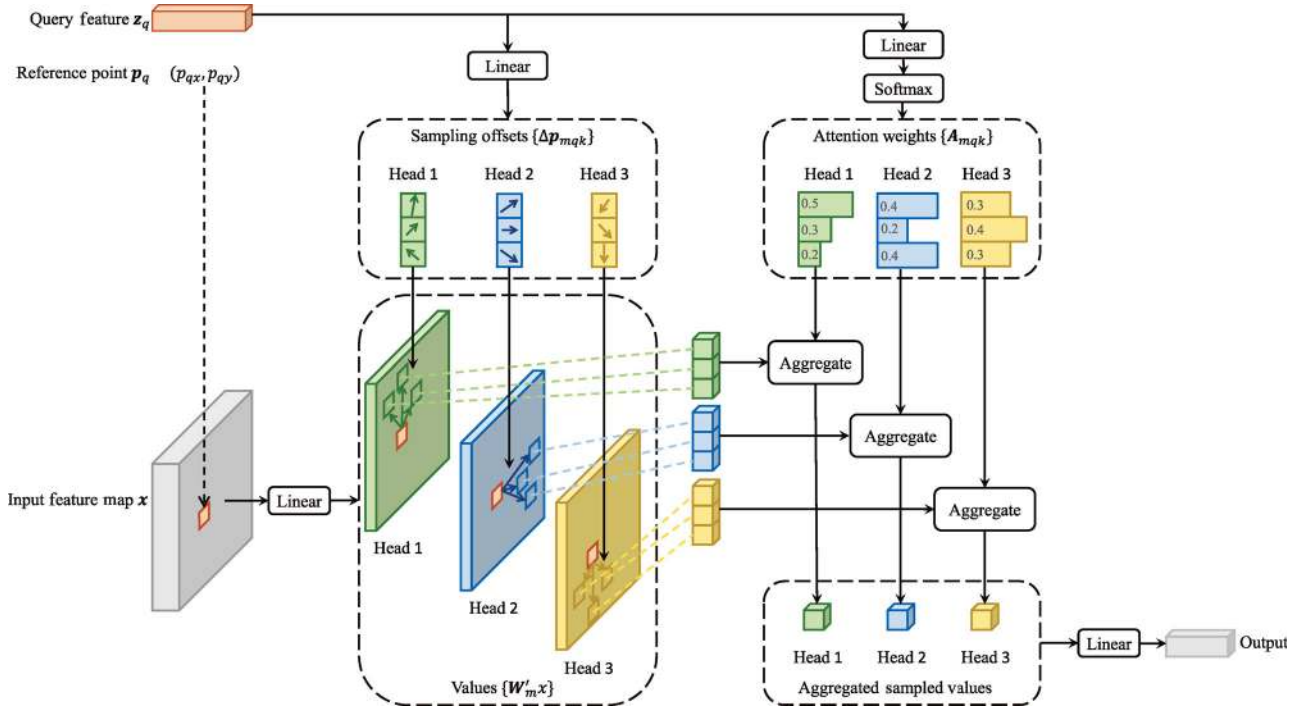

Fig. 10 Deformable attention module. Reproduced with permission from Ref. [25], © The Author(s) 2020.

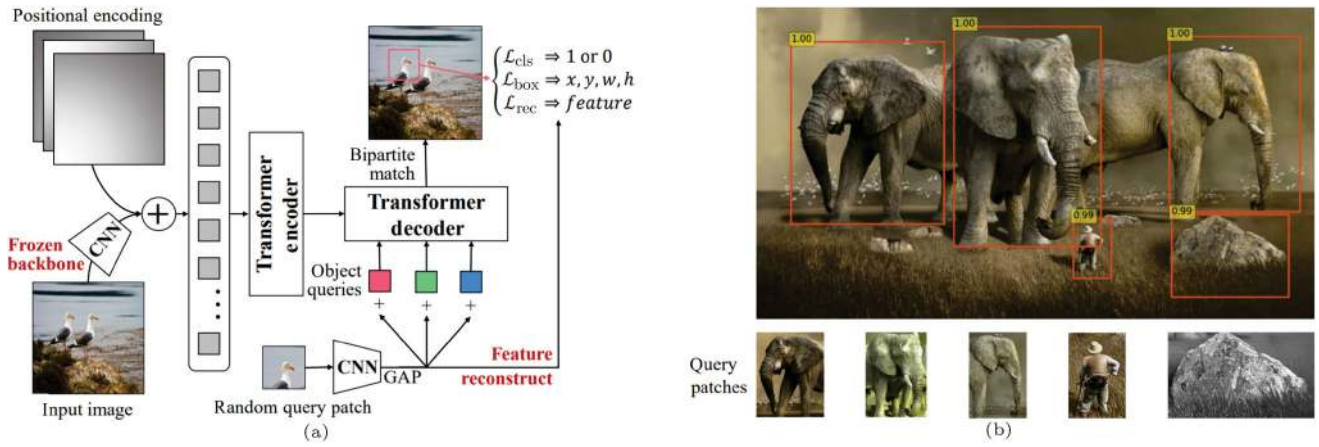

Fig. 11 (a) Random single query patch detection of UP-DETR. (b) More robust representation is derived by augmenting the query patch. Reproduced with permission from Ref. [26], © The Author(s) 2020.

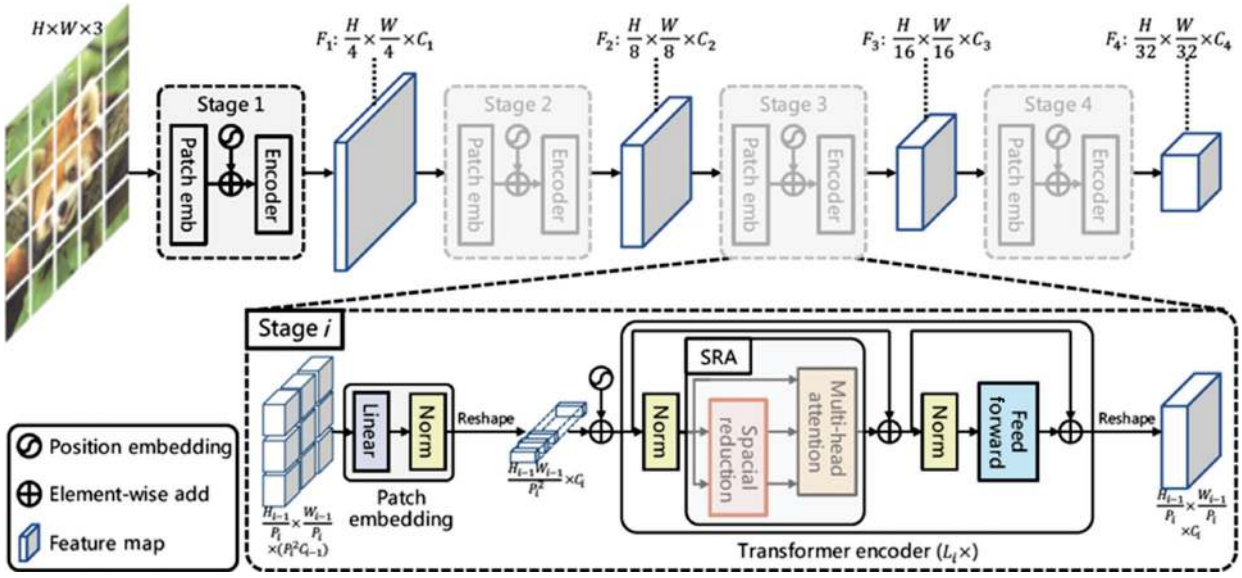

Fig. 12 Framework of PVT. Reproduced with permission from Ref. [27], © The Author(s) 2021.

tasks, as listed in Table 4. Low-level vision tasks include super-resolution [101], denoising [101], image colorisation [30], text-to-image generation [31], and image generation [34, 35]. We separately introduce how these tasks use transformers to achieve good results (see examples in Fig. 13).

Table 4 Source code links for ViT-based models for low-level vision tasks

Method	Source (GitHub)
TIME [31]	—
IPT [101]	—
ColTran [30]	`google-research/google-research/tree/master/coltran`
TTSR [73]	`researchmm/TTSR`
GANsformer [35]	—
TransGAN [34]	`VITA-Group/TransGAN`
DALL·E [32]	`openai/DALL-E`
VQGAN [102]	`CompVis/taming-transformers`
StyTr^2 [38]	—
PCT [39]	`Strawberry-Eat-Mango/PCT_Pytorch`

5.1 TIME

As a pre-trained NLP model is always required for the text-to-image (T2I) task, it may introduce inflexibility for the whole model. Liu et al. [31] propose an efficient model for T2I tasks: Text and Image Mutual Translation Adversarial Networks (TIME). TIME can jointly handle T2I and image captioning using a single network without a pretrained NLP model. As Fig. 14 shows, TIME introduces a multi-head and multi-layer transformer to the generator and text decoder, which can be used to effectively

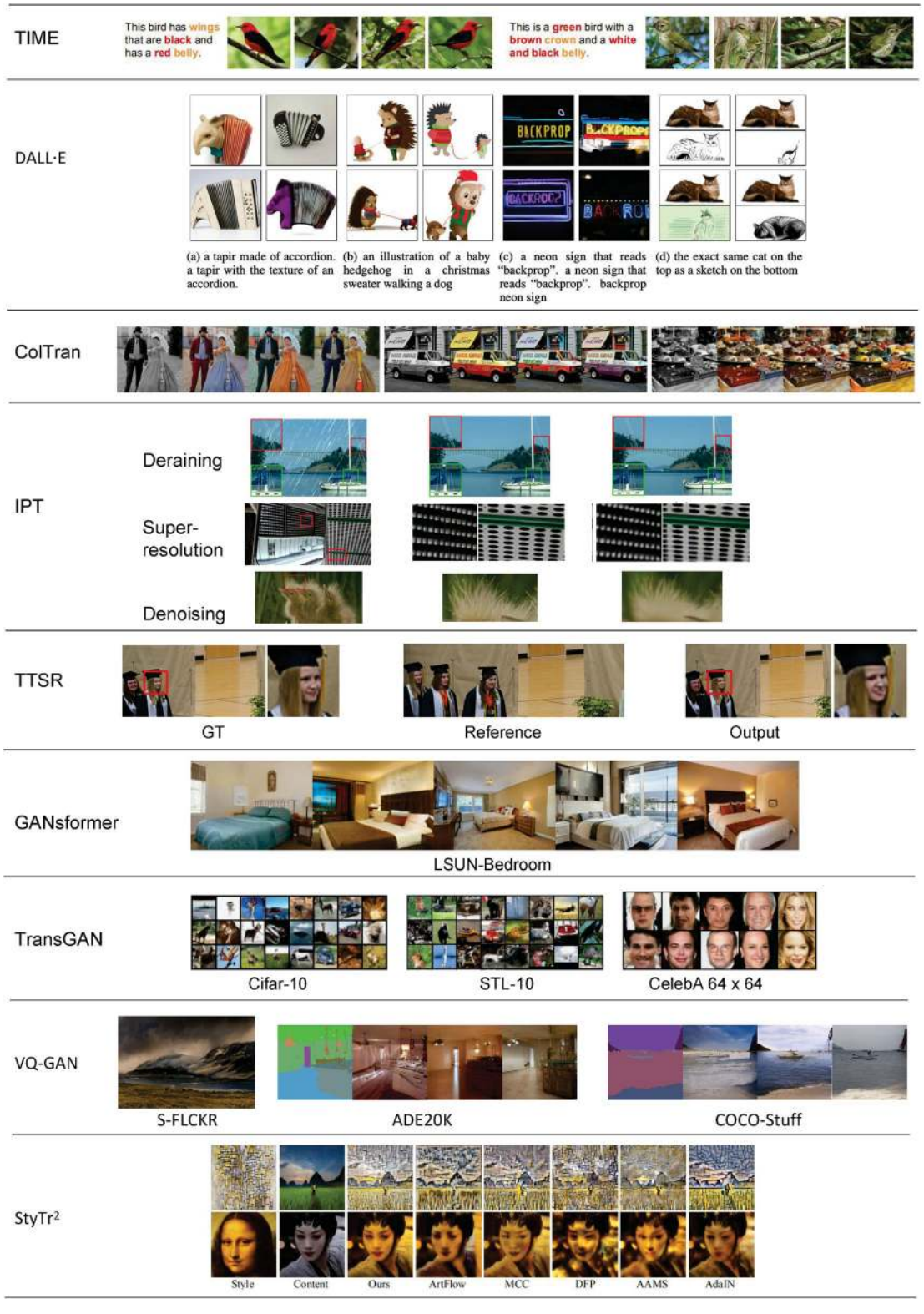

Fig. 13 Representative results for low-level tasks, such as text-to-image generation, basic image processing tasks, colorization, image super resolution, and image generation. Images are taken from the corresponding papers.

combine image features and the sequence of word embeddings into the output. The Text-Conditioned Image Transformer takes image feature f_i and the sequence of word embeddings f_t, and outputs the revised image f_{it} according to the word embeddings. The Image-Captioning Transformer is similar to the

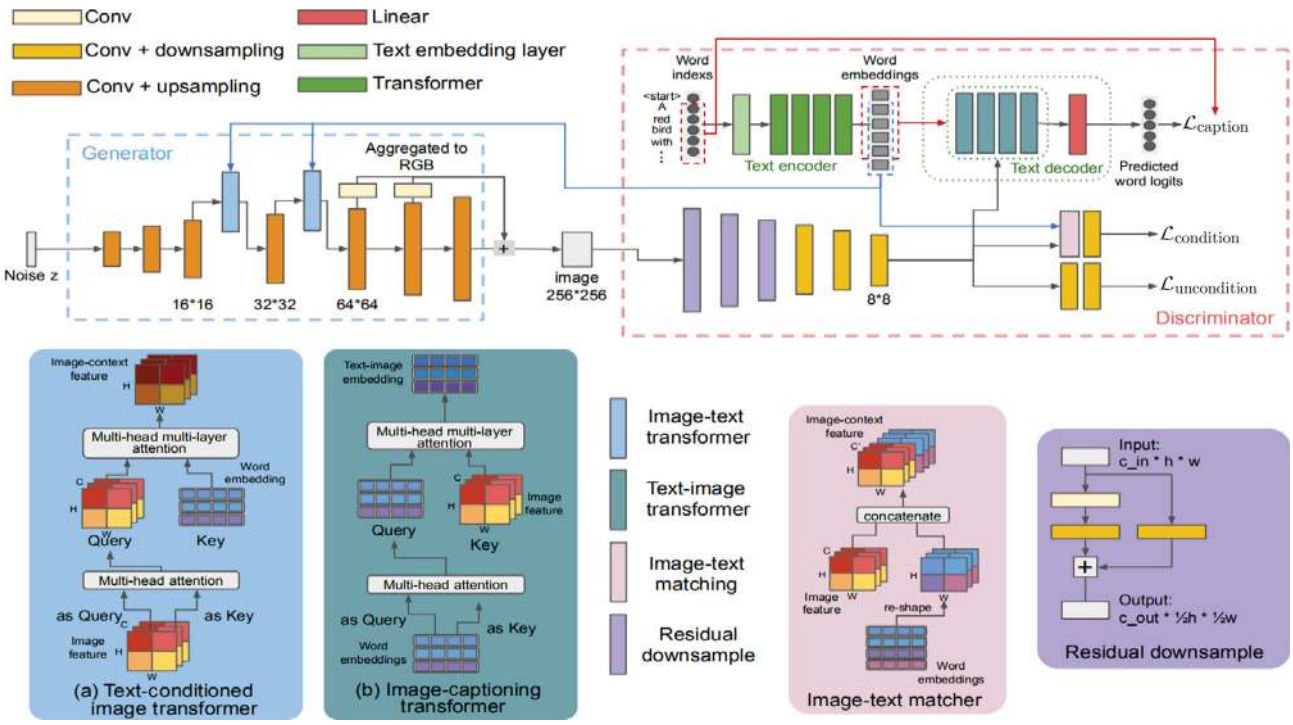

Fig. 14 TIME model overview. Reproduced with permission from Ref. [31], © Association for the Advancement of Artificial Intelligence 2021.

Text-Conditioned Image Transformer but for image captioning [103–105]. T2I and the image captioning task are jointly trained in the generative adversarial network (GAN) manner. TIME achieves state-of-the-art T2I performance without pretraining.

DALL·E. Text-to-image generation is a classical generation problem, which needs to construct a mapping between two streams. Ramesh et al. [32] propose a transformer-based framework to better align text and image semantic information. A two-stage model is applied to model the text and image tokens. They first train a discrete variational autoencoder [106] to build 1024 image tokens and adopt 256 BPE-encoded text tokens to represent the text information. Thereafter, an auto-regressive transformer is used to capture the joint distribution of the text and image tokens. They also use a mixed-precision training strategy and PowerSGD [57] to save GPU memory. The model consumes approximately 24 GB memory in 16-bit precision.

5.2 IPT

Classification models can be pretrained on large-scale datasets to enlarge model representation ability. Related low-level vision tasks such as image super-resolution, inpainting, and deraining are combined in a model to help one another. The generalized pretraining procedure solves the problem of task-specific data limitation. Therefore, Chen et al. [101] develop a pretrained model for image processing using the transformer architecture, the Image Processing Transformer (IPT). The model architecture is shown in Fig. 15. To adapt to different vision tasks, Chen et al. [101] design a multi-head and multi-tail architecture, which involves three convolutional layers. The transformer body consists of an encoder and a decoder described in Ref. [57]. Like the discriminator in Ref. [34], they split the given features into patches and each patch is regarded as a "word" before features are input into the transformer body. Unlike the original transformer, they utilize a task-specific embedding as an additional input to the decoder. The model is pretrained on ImageNet, which is a key factor for success.

5.3 Uformer

Wang et al. [37] propose an effective and efficient transformer-based architecture for image restoration. It uses a transformer module to construct a hierarchical encoder–decoder network. Two core designs of Uformer make it suitable for image restoration. The first is a local-enhanced window transformer block. Specifically, a nonoverlapping window-based self-attention is used to reduce the

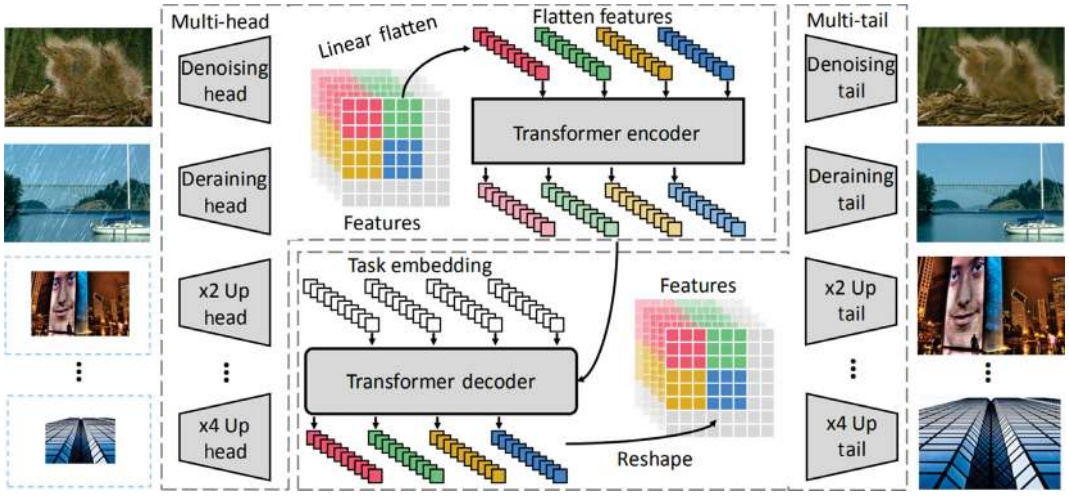

Fig. 15 Overview of IPT. Reproduced with permission from Ref. [101], © The Author(s) 2021.

computational cost, and depth-wise convolution is used in the FFN to further improve its ability to capture local context. The second is the skip-connection mechanism, which is explored to effectively deliver the encoder information to the decoder. Uformer can capture useful dependencies for image restoration because of the two designs above. The network structure of Uformer is shown in Fig. 16. Its performance has been verified through several image restoration tasks, including denoising, deraining, and deblurring.

5.4 TransGAN

Driven by curiosity, Jiang et al. [34] first design a GAN using pure transformer-based structures to determine whether transformers perform well when applied to generative adversarial networks (GANs) [107]. This network consists of a memory-friendly transformer-based generator and a patch-level discriminator. Jiang et al. [34] also imitate the philosophy in CNN-based GANs and design a novel structure for image generation to avoid the high cost when applying transformers from NLP

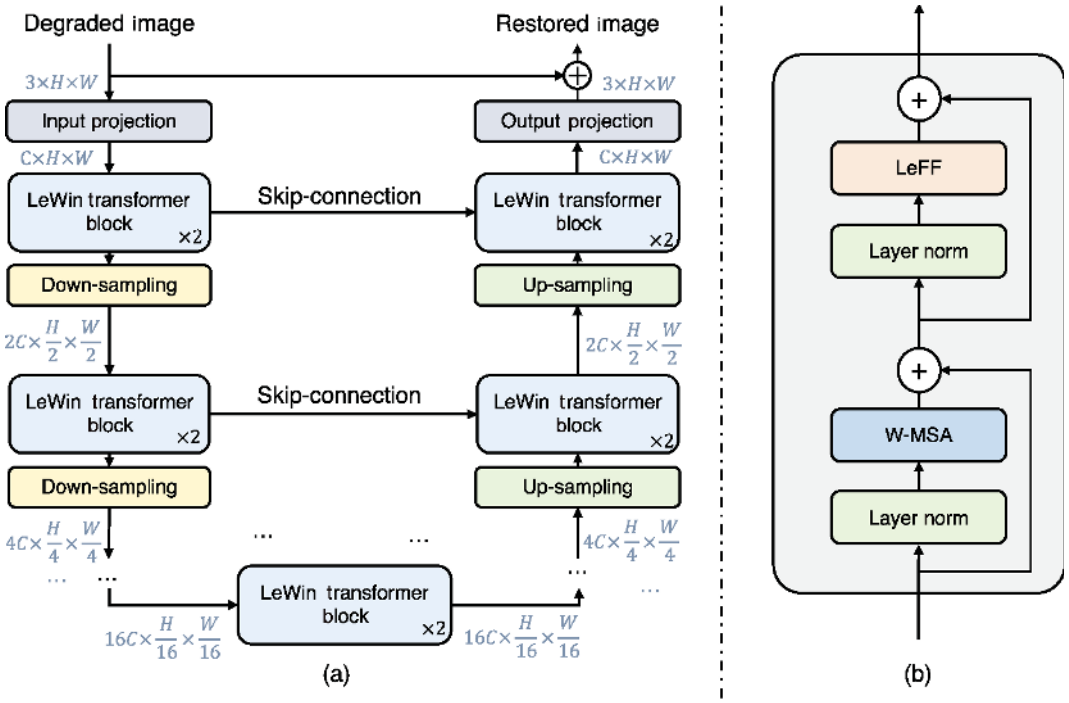

Fig. 16 (a) Overview of Uformer. (b) Structure of the LeWin transformer block. Reproduced with permission from Ref. [37], © The Author(s) 2021.

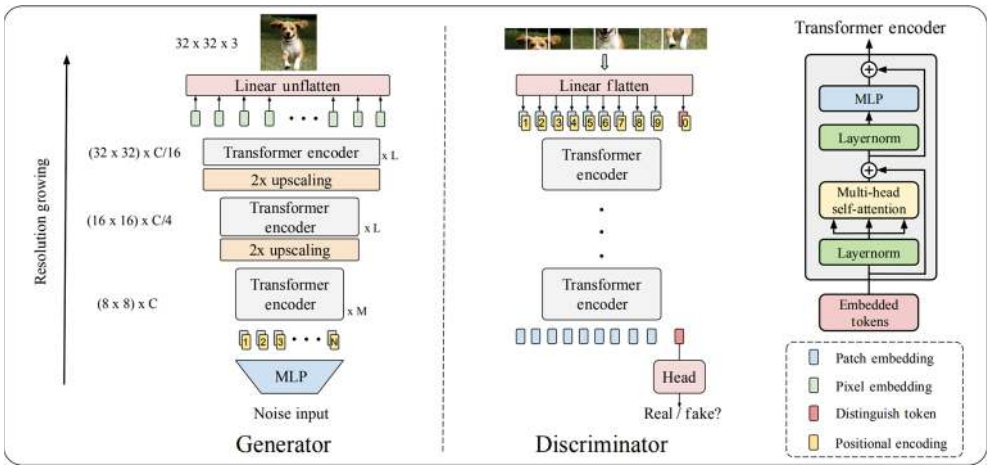

Fig. 17 Model overview of TransGAN. Reproduced with permission from Ref. [34], © The Author(s) 2021.

to visual tasks. As shown in Fig. 17(left), the memory-friendly transformer-based generator has multiple stages, thus increasing the feature resolution while decreasing the embedding dimension. The discriminator splits the generated images into small patches and regards them as "words". The tokens are taken by the classification head to output the real/fake prediction. The whole net is trained with three ingenious strategies: data augmentation, self-supervised auxiliary task (super task) cooperative training, and locality-aware initialization. The results in CIFAR-10 and STL-10 are comparable to those of some state-of-the-art works using CNN-based GANs.

5.5 TTSR

Texture is often damaged during downsampling and also cannot be easily recovered. Traditional single image super-resolution always leads to blurring effects in the output. Therefore, Yang et al. [33] propose a reference-based image super resolution method, namely, the Texture Transformer Network for Image Super Resolution (TTSR). As shown in Fig. 18, the Learnable Texture Extractor is first used to extract proper texture information, which is crucial for super resolution. Then, the input to the texture transformer can be expressed as follows:

$$Q = \mathrm{LTE}(\mathrm{LR}\uparrow)$$
$$K = \mathrm{LTE}(\mathrm{Ref}\downarrow\uparrow)$$
$$V = \mathrm{LTE}(\mathrm{Ref})$$

where LR↑, Ref, and Ref↓↑ denote the image to be reconstructed, the reference image, and the reference image that is down-sampled and then up-sampled respectively. The texture transformer contains a

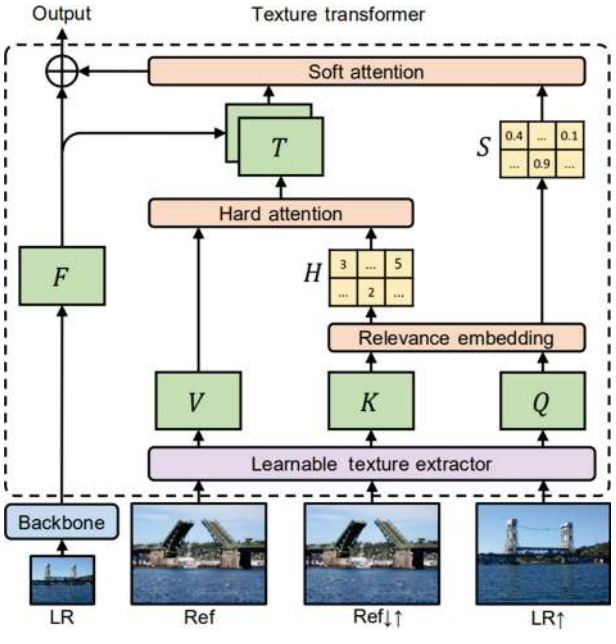

Fig. 18 Model overview of TTSR. Reproduced with permission from Ref. [33], © IEEE 2020.

Hard-Attention and a Soft-Attention, and it is applied to the high-resolution feature guided by the reference image. Finally, they propose a cross-scale feature integration module to exchange information between the features at different scales for better representation at different scales.

5.6 ColTran

Image colorization is a challenging task that needs to determine the image semantics. Most colorization models estimate log-likelihood based on neural generative approaches. Kumar et al. [30] propose the Colorization Transformer (ColTran) using a

self-attention mechanism to promote the effects of a probabilistic colorization model. ColTran replaces self-attention blocks with axial self-attention, which decreases the computational complexity from $\mathcal{O}(D^2)$ to $\mathcal{O}(D\sqrt{D})$. Kumar et al. [30] adopt a conditional variant of the Axial Transformer [108] for low-resolution coarse colorization. As shown in Fig. 19, the ColTran core consists of Conditional Self-Attention, MLP, and Layer Norm modules, and it applies conditioning to the auto-regressive core. They also design a Color Upsampler and Spatial Upsampler to produce high-fidelity colorized images from low resolution results. The Color Upsampler converts the coarse image of 512 colors back into a 3-bit RGB image with 8 symbols per channel. The Spatial Upsampler generates colorized images with high resolution. ColTran can handle grayscale images of 256×256 pixels.

5.7 GANsformer

The cognitive science literature talks about two mechanisms by which human perception interacts, namely, bottom–up and top–down processing. Previous vision tasks using CNNs do not reflect this bidirectional nature because the local receptive field reduces their ability to model long-range dependencies. Therefore, Hudson and Zitnick [35] aim to design a transformer network with a highly adaptive architecture centered around relational attention and dynamic interaction. They propose a Bipartite Transformer to eliminate the limitation of huge computational complexity of self-attention of transformers. Unlike the self-attention operator which considers all pairwise relations between input elements, the Bipartite Transformer generalizes this formulation by featuring a bipartite graph between two groups of variables (latent and image features) instead. As shown in Fig. 20, simplex attention distributes information in a single direction over the Bipartite Transformer, while Duplex attention supports bidirectional interaction between the elements. The bipartite structure makes a good balance between expressiveness and efficiency, and it constructs the interaction between latent and visual features to generate good results.

5.8 StyTr^2

Considering the limited receptive fields of CNNs, obtaining global information about input images is difficult but is critical for the image style transfer task. The content leak problem also occurs when CNN-based models are adopted for style transfer.

Therefore, Deng et al. [38] propose the first transformer-based style transfer model using the ability for long-range extraction (Fig. 21). The unbiased Style Transfer Transformer framework StyTr^2 contains two transformer encoders to obtain domain-specific

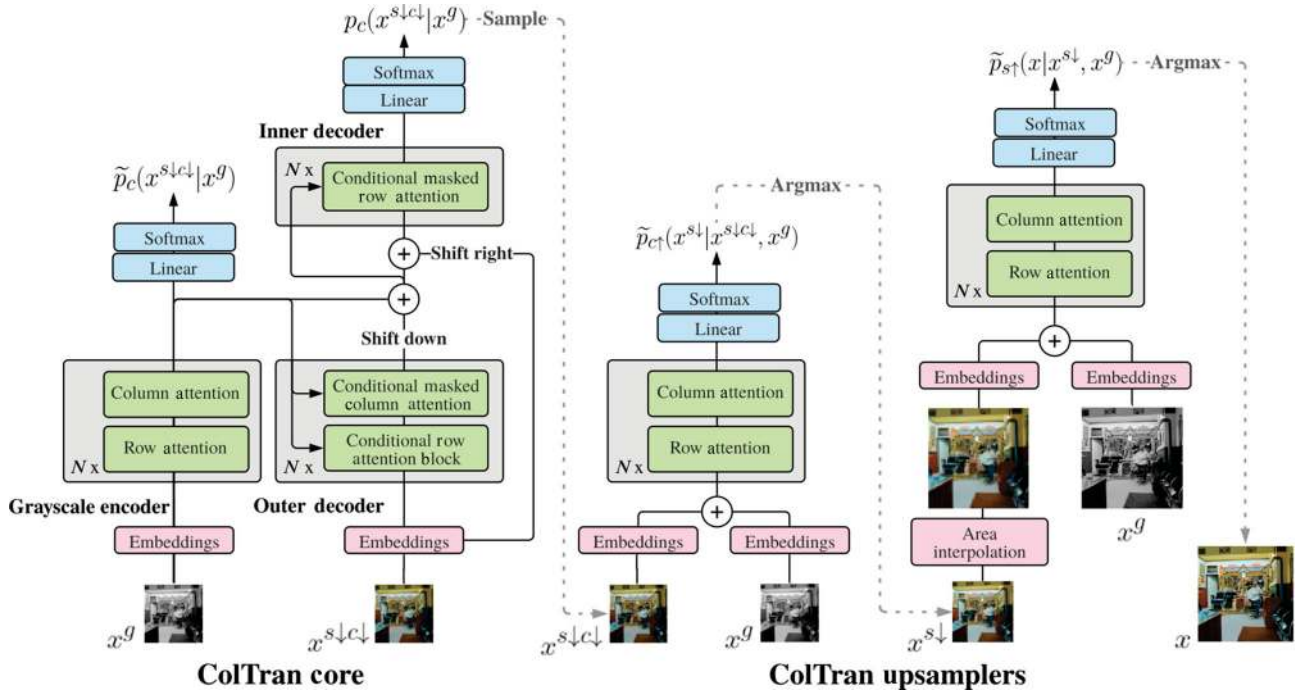

Fig. 19 Overview of colorization transformer. Reproduced with permission from Ref. [30], © The Author(s) 2021.

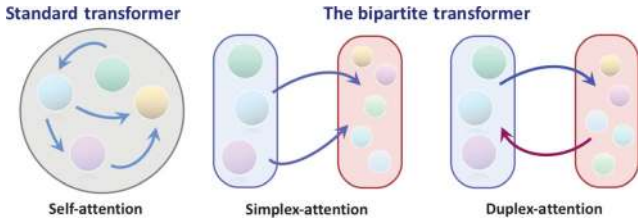

Fig. 20 Overview of the GANsformer framework. Reproduced with permission from Ref. [35], © The Author(s) 2021.

information. Following encoding, a multilayer transformer decoder generates the output sequences. Moreover, Deng et al. [38] propose a content-aware mechanism to learn the positional encoding based on image semantic features and dynamically expand the position to suit different image sizes.

5.9 VQGAN

High-resolution image synthesis is a difficult generation problem which aims to generate high-fidelity images within a reasonable time. Convolutional approaches exploit the local structure of the image, while transformer methods are good at establishing long-range interactions. Esser et al. [102] utilize the advantages of CNNs and transformers to build a high-resolution image generation framework. They propose a variant of VQVAE [36] and adopt adversarial learning to achieve vivid results. The content hidden space consists of a discrete codebook, and different codes in the codebook are combined according to a certain probability to represent the content information. The key to sampling in a discrete space is to predict the distribution of discrete codes, and the transformer can deal with the issue. Given the first i codes, the transformer module is used to predict the probability of occurrence of the i-th code. The number of codes in the codebook is 512–4096 according to the dataset. The model can synthesize the results containing 1280×460 pixels.

5.10 PCT

Unlike CNNs, transformers are inherently permutation invariant when processing a series of points and are thus suitable for point cloud learning. Guo et al. [39] propose a state-of-the-art transformer-based point cloud model based on offset-attention with an implicit Laplace operator. They enhance the input embedding based on farthest point sampling and nearest neighbor search to better capture the local context in the point cloud.

6 Multimodal learning

The above sections cover developments in conventional computer vision. Apart from pure vision tasks, transformer-based models have also achieved promising progress in language and vision multimodal tasks, such as visual question answering (VQA) [109, 110], image captioning [111], and image retrieval [112], due to the high performance achieved by the NLP transformers. Transformer-based vision-language (V+L) approaches are often pretrained on multiple tasks and fine-tuned on diverse downstream sub-tasks. Inputs of different modalities share the analogous single- or two-stream architecture.

In this section, we start from recently representative transformer-based works on V+L tasks with different frameworks (Section 6.1), and then summarise pretraining objectives (Section 6.2) and compare details (Section 6.3).

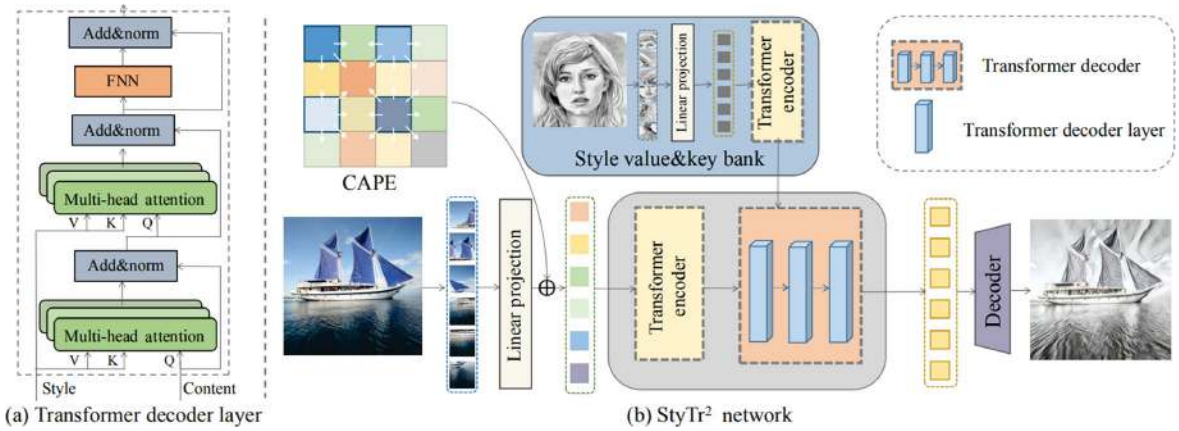

Fig. 21 Overview of StyTr^2. Reproduced with permission from Ref. [38], © The Author(s) 2021.

6.1 Transformer-based V+L works

Most transformer-based V+L works are based on two kinds of structures: the two-stream (each stream for a single modality) framework or the single-stream (common stream for jointly learning cross-modal representation) framework. ViLBERT [40] and UNITER [41] are representative works for two- and single-stream frameworks, respectively. Meanwhile, SemVLP [42] unifies the two mainstream architectures for aligning the cross-modal semantics.

6.1.1 ViLBERT

ViLBERT [40] is a representative two-stream transformer-based model for V+L. Two separate streams are used for vision and language processing. Figure 23 shows the architecture of ViLBERT. Two parallel BERT-style models operate on image regions and text tokens. Each stream connects a series of transformer blocks (TRM) and co-attentional transformer layers (Co-TRM). As shown in Fig. 22, the Co-TRM layers enable information exchange between modalities, and the modified attention mechanism is the key technical innovation. By exchanging key–value pairs in multi-headed attention, the Co-TRM structure allows for variable network depth for each modality and enables cross-modal connections at different depths.

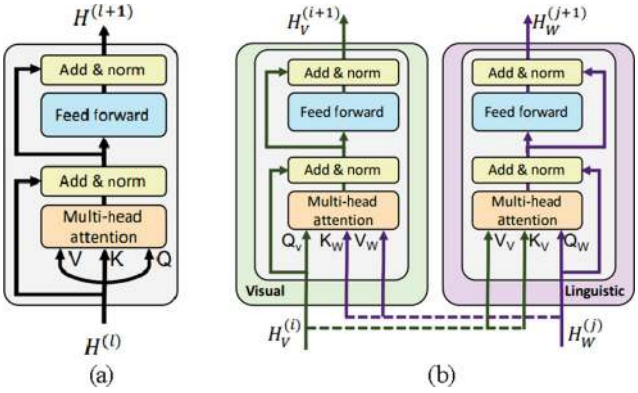

Fig. 22 (a) Architecture of a standard encoder transformer block. (b) Co-attention transformer layer in ViLBERT. Reproduced with permission from Ref. [40], © The Author(s) 2019.

6.1.2 UNITER

Chen et al. [41] propose UNITER: UNiversal Image-TExt Representation. It can power heterogeneous downstream V+L tasks with joint multimodal embeddings. As shown in Fig. 24, UNITER first encodes image regions (visual features and bounding box features) and textual words (tokens and positions) into a shared embedding space with image and text embedders. Then, UNITER applies a transformer module to learn the joint embedding of the two modalities through designed pretraining tasks that include classic image–text matching (ITM), masked language modeling (MLM), and masked region modeling (MRM). UNITER uses conditional masking on MLM and MRM, which means masking only one modality while keeping the other untainted. A novel word–region alignment pretraining task via optimal transport is also proposed to encourage fine-grained alignment between words and image regions. The authors consider the matching of word tokens and RoI regions as minimizing the distance of two discrete distributions, where the distance is computed based on optimal transport. UNITER, as a single-stream model, achieved state-of-the-art performance when proposed. ViLLA [113], which combines UNITER and adversarial training, achieves higher performance.

6.1.3 SemVLP

Li et al. [42] present a novel V+L framework, SemVLP. It unifies both mainstream architectures. By fusing single- and two-stream architectures, SemVLP utilizes cross-modal semantics. Its framework is detailed in Fig. 25. On the basis of a shared bidirectional transformer encoder with cross-modal attention module, SemVLP can encode the input text and image into different semantics. It adopts common pretraining methods with a special training strategy: single- and two-stream frameworks are updated in

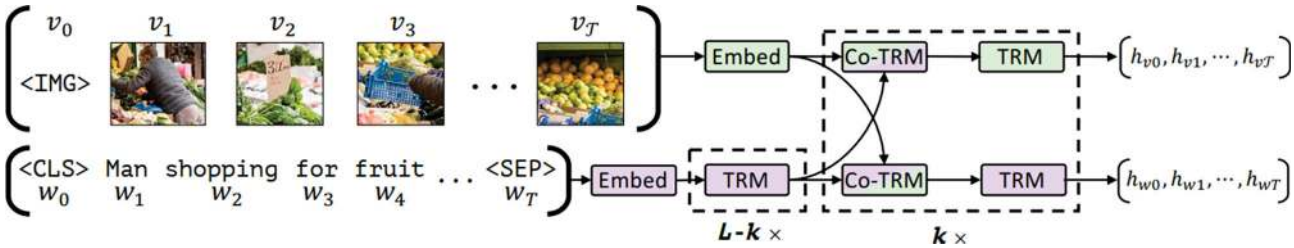

Fig. 23 Overview of ViLBERT. Reproduced with permission from Ref. [40], © The Author(s) 2019.

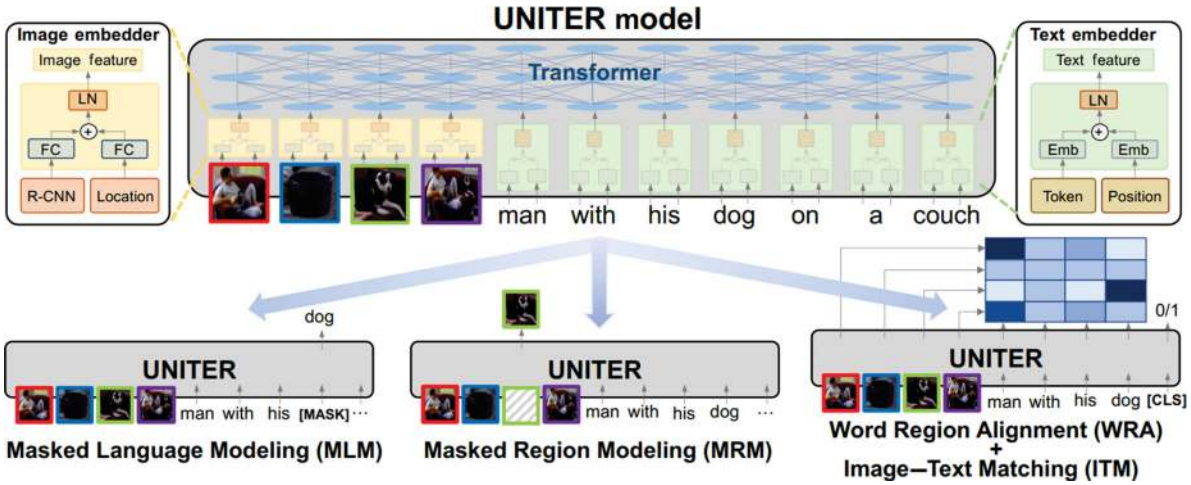

Fig. 24 Overview of UNITER. Reproduced with permission from Ref. [41], © Springer Nature Switzerland AG 2020.

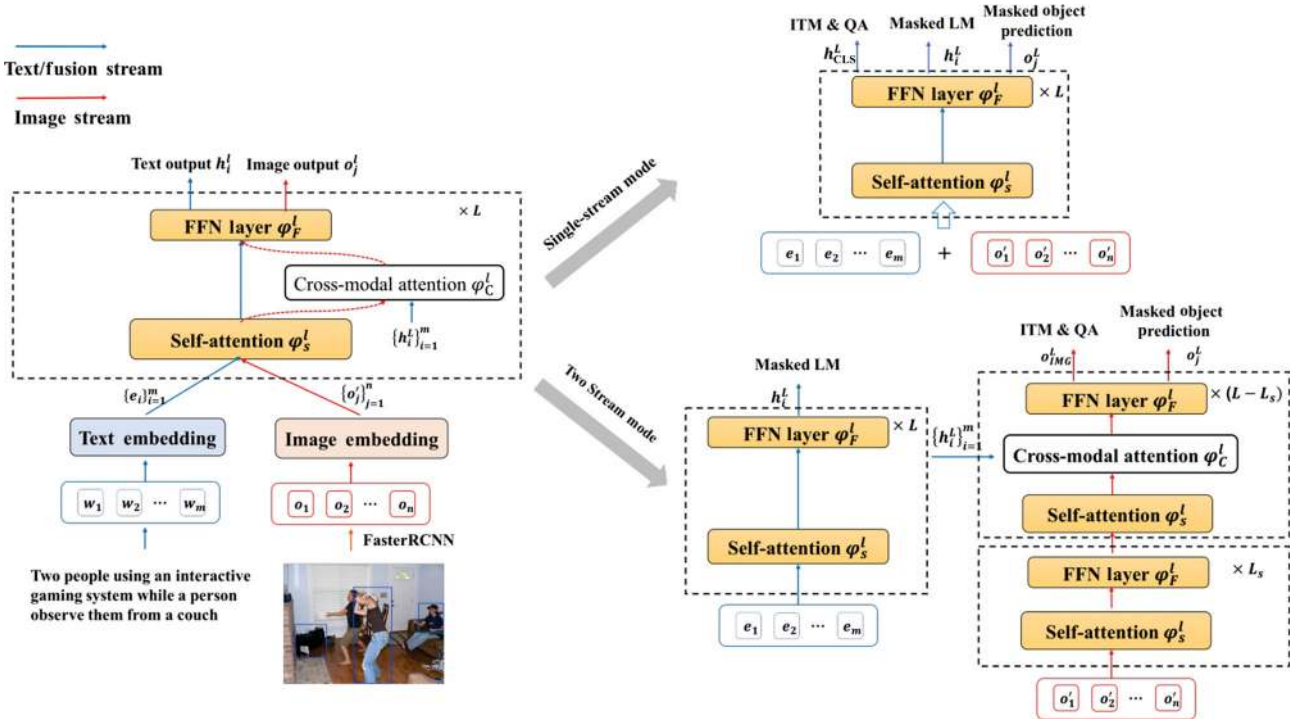

Fig. 25 Overview of SemVLP. Reproduced with permission from Ref. [42], © The Author(s) 2021.

each half of the training time for each mini-batch of image–text pairs.

6.2 Multimodal pretraining

Designing reasonable pretraining objectives for transformer-based models, such as masked language modeling (MLM) and next sentence classification from BERT, has brought excellent results on NLP tasks. These methods also work in the cross-modal field with V+L. The key challenge is the way to replicate or extend large-scale pretraining to cross-modal methods and to design novel pretraining objectives for multimodal learning. In this section, we briefly introduce pretraining tasks extended from BERT. These extended approaches include MLM, masked region modeling (MRM), and image–text matching (ITM). We also list other specially designed pretraining tasks for multimodal learning.

6.2.1 Masked language modeling

Most recent V+L works follow BERT in using MLM for cross-modal tasks. UNITER modifies MLM by introducing visual information. Specifically, UNITER attempts to predict masked words based on

observation of the surrounding words and all image regions. InterBERT [114] changes MLM to masked segment modeling. In the case of using a random word to replace the selected word, masked segment modeling masks a continuous segment of text instead of random words.

6.2.2 Image–text matching

For another pretraining task of BERT, next sentence classification has been converted to an ITM problem, which determines whether a pair of sentence and image regions match. This task is widely used in advanced V+L works. InterBERT [114] performs ITM with hard negatives by regarding the image–text pairs in the dataset as positive samples, pairing the images with uncorrelated texts, and regarding the pairs as negative samples. VL-BERT [115] and UnifiedVLP [116] also do not use ITM, tending to use other efficient choices like MRM introduced next.

6.2.3 Masked region modeling

The existing masking method, MRM, is the dual task of MLM. MLM can be easily applied to visual input. Some researchers have proposed several novel pretraining methods by masking input visual tokens to extend masked modeling to vision. Masked region feature regression (MRFR) is one of these approaches applied by ViLBERT [40]. ViLBERT trains the model to regress the masked input RoI pooled feature, which is extracted by Faster R-CNN [86]. Most models perform optimization with L2 loss. VL-BERT [115] also follows MRFR instead of using ITM. It uses masked RoI classification with linguistic clues, predicting the category label of the masked RoI obtained by Fast R-CNN [117] from the other clues. On the contrary, some models choose masked region classification, which lets the model predict the object semantic class for each masked region. Models are often optimized by cross-entropy loss or KL-divergence to learn the class distribution. These MRM tasks are performed in UNITER and UNIMO [118]. InterBERT [114] also changes MRM strategy in the visual modality by masking objects which have a high proportion of mutual intersection with zero vectors to avoid information leakage due to overlap between objects. Notably, earlier transformer-based works, such as VisualBERT [119] and B2T2 [120], do not extend MLM to the visual domain.

6.2.4 Other designs for V+L

Some models are also trained with unique, newly designed pretraining strategies. In Oscar [121], each image–text pair is defined as a triple and thus consists of a word sequence, a set of object tags, and a set of image region features. Therefore, in addition to MLM on words and object tags, Oscar uses a contrastive loss to encourage the model to distinguish the original and modified triple. By differently using contrastive learning, UNIMO creates image–text pairs by a novel text rewriting method. ERNIE-ViL [122] introduces a scene graph to design advanced pretrained tasks, including object prediction, attribute prediction, and relationship prediction. Li et al. [123] add masked sentence generation to optimize their model: a cross-modal decoder is taught to autoregressively decode the input sentence word-by-word conditioned on the input image. Training directly on downstream tasks like QA is also used in LXMERT [124] and SemVLP [42].

6.3 Comparisons and implementation details

Table 5 details implementations and open source. The MSCOCO dataset [111], maintained by Microsoft, is widely used in multiple tasks like object detection. The Conceptual Captions Dataset (CC) [125] is provided by Google AI and consists of nearly 3.3 million images annotated with captions harvested from the web. The SBU Captions Dataset [126] includes image captions collected from 1 million images from Flickr[①]. MSCOCO, CC, and SBU all can be used for image caption tasks. The Visual Genome [127] is a dataset including images and image content semantic information. Visual Genome, VQA [109], VQAv2 [110], and GQA [128] datasets can all be used for VQA pretraining. Notably, partial datasets are used as benchmarks simultaneously. Table 6 shows the performance of models reported above on different V+L benchmark datasets. The results are obtained by models fine-tuned on the corresponding datasets.

7 Conclusions and discussion

7.1 Backbone design

Section 3 describes several recent developments in the backbone design of visual transformers, including

① https://www.flickr.com

Table 5 Model setting in various papers. COCO refers to MS COCO [129], CC to Conceptual Captions [125], VG to Visual Genome [127], SBU to SBU captions [126], and OI to OpenImages [130]

Model	Dataset(s) for pre-training	Params	Batch size	Hard-aware	Source (GitHub)
ViLBERT [40]	CC	221M	512	8 TitanX	`jiasenlu/vilbert_beta`
VL-BERT [115]	CC	110M	256	16 V100	`jackroos/VL-BERT`
UNITER [41]	CC/COCO/VG/SBU	110M	Dynamic	16 V100	`ChenRocks/UNITER`
Oscar [121]	CC/COCO/VG/SBU/Flicker30K/VQA/GQA	110M	512	—	`microsoft/Oscar`
VILLA [113]	CC/COCO/VG/SBU	—	Task-specific	—	`zhegan27/VILLA`
ERNIE-ViL [122]	CC/SBU	210M	512	8 V100	`PaddlePaddle/ERNIE`
UNIMO [118]	CC/COCO/VG/SBU	—	—	—	`weili-baidu/UNIMO`
VinVL [131]	CC/COCO/VG/SBU/Flicker30K/VQA/GQA/OI	—	1024	—	`pzzhang/VinVL`
TDEN [123]	CC	—	1024	16 P40	`YehLi/TDEN`
UniT [132]	COCO/VG/VQAv2	—	64	64 V100	—
SemVLP [42]	CC/COCO/VG/SBU/VQAv2/GQA	140M	256	4 V100	—

Table 6 Comparison of transformer-based V+L models on VQA [109], GQA [128], Flickr30K [112], CoCo Caption [111], NLVR2 [133], SNLI-VE [134], VCR [135], and RefCOCO+ [136] benchmarks

	VQA		GQA		IR-Flickr30K			TR-Flickr30K			CoCo Caption		NLVR2		SNLI-VE		VCR			RefCOCO+		
	test-dev	test-std	test-dev	test-std	R@1	R@5	R@10	R@1	R@5	R@10	BLUE4	CIDEr	dev	test-P	val	test	Q/A	QA/R	Q/AR	val	testA	testB
ViLBERT [40]	70.55	70.92	—	—	58.20	84.90	91.52	—	—	—	—	—	—	—	—	—	73.3	74.6	54.8	72.34	78.52	58.20
VL-BERT [115]	71.79	72.22	—	—	—	—	—	—	—	—	—	—	—	—	—	—	75.8	78.4	59.7	80.31	83.62	75.45
UNITER [41]	73.82	74.02	—	—	75.56	94.08	96.76	87.3	98.0	99.2	—	—	79.12	79.98	79.39	79.38	77.3	80.8	62.8	84.25	86.34	79.75
Oscar [121]	73.61	73.82	61.58	61.62	—	—	—	—	—	—	41.7	140.0	79.12	80.37	—	—	—	—	—	84.40	86.22	80.00
VILLA [113]	74.69	74.87			76.26	94.24	96.84	87.9	97.5	98.8	—	—	79.76	81.47	80.18	80.02	78.9	79.1	60.6	84.40	86.22	80.00
ERNIE-ViL [122]	74.95	75.10	—	—	76.66	94.16	96.76	89.2	98.5	99.2	—	—	—	—	—	—	79.2	83.5	66.3	75.89	82.37	66.91
UNIMO [118]	73.79	74.02	—	—	—	—	—	—	—	—	38.6	124.1	—	—	80.00	79.10	—	—	—	—	—	—
VinVL [131]	76.52	76.60	65.05	64.65	75.40	92.90	93.30	58.8	83.5	90.3	41.0	140.9	82.67	83.98	—	—	—	—	—	—	—	—
TDEN [123]	72.50	72.80	—	—	—	—	—	—	—	—	40.2	133.4	—	—	—	—	75.7	76.4	58.0	—	—	—
SemVLP [42]	74.52	74.68	62.87	63.62	74.80	93.43	96.12	87.7	98.2	99.3	—	—	79.00	79.55	—	—	—	—	—	—	—	—

feature map visualization approaches. Recent progress can be technically divided into two main streams: (i) enhancing the capability of visual transformers in modeling spatial structure and locality mechanism, such as a better image-to-token module, a pixel-level transformer block, a depth-wise convolution-based pooling layer, and an SW-MSA module, and (ii) boosting the richness of learned visual features and promoting efficient use of parameters, such as conditional position encoding, a message communication scheme between the MSA heads, and deep-narrow ViT architectures.

As the first visual transformer was proposed very recently (October 2020), we believe that the potential of the ViT model has not been fully exploited and several research topics are worthy of consideration and effort:

- Advanced designs of basic ViT operation or modules and the corresponding learning scheme for CV tasks, like injecting prior knowledge of image data or the computer vision task into the module design or the learning scheme of visual transformer models, and making the transformer more computationally efficient, are of interest. The versatility of ViT models in additional real-world scenarios, such as aesthetic visual analysis [137–139], face anti-spoofing [140, 141], and point cloud learning [142], is also worthy of exploration.
- The transformer block can be placed in the perspective of NAS. One of the goals of the NAS framework [143–145] is to search for optimal network architectures for a given task without human intervention. Interesting architectures can be considered and practical insights for further developments can be gained by building on a well-designed search space that contains a transformer block. Several recent works have investigated this topic. Wang et al. [146] and So et al. [147] leverage NAS techniques to seek for effective and efficient transformer-based architectures automatically. Li et al. [148] propose a novel scheme, BossNAS, to achieve optimal solutions which trade-off CNN architecture and transformer blocks.

- Understanding of the working mechanism and theoretical rationale of visual transformers can be enhanced. Several researchers have achieved promising progress in unveiling the power of transformer models, from such perspectives as information bottlenecks [149, 150] and better visualization tools [23, 63].

7.2 High-level vision

In Section 4, we introduce several representative works on object detection. The basic logic follows the line of DETR [24]. PVT [27], which is a general backbone for dense prediction, is also introduced. Several problems still need to be addressed despite improvements brought by these works. Unlike CNN-based methods, such as Faster-RCNN [86], current transformers for dense prediction tasks suffer from high computation time. Thus, efficiency of transformers for high-level vision remains a pressing research direction.

7.3 Low-level vision and generation

In Section 5, we introduce some low-level vision and image generation tasks using transformer-based models. They can achieve outstanding results but have difficulty generating large images. Therefore, extending a pure transformer with CNN layers is widely adopted by many works. A pure transformer structure still faces the challenge of high computation time.

7.4 Multimodal learning

In Section 6, we introduce several representative transformer-based models proposed in the past 2 years for vision and language tasks. We also review mainstream pretraining tasks in the V+L field. Meanwhile, transformer-based models have succeeded for the tasks listed in Table 6, but performance can still be improved:

- Pure transformers may be an alternative choice for the image mode.
- Design of efficient pretraining tasks can lead to better results and performance.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported by National Key R&D Program of China under Grant No. 2020AAA0106200, and by National Natural Science Foundation of China under Grant Nos. 61832016 and U20B2070.

References

[1] He, K. M.; Zhang, X. Y.; Ren, S. Q.; Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.

[2] Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, 2019.

[3] Radosavovic, I.; Kosaraju, R. P.; Girshick, R.; He, K. M.; Dollár, P. Designing network design spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10425–10433, 2020.

[4] Yin, M. H.; Yao, Z. L.; Cao, Y.; Li, X.; Zhang, Z.; Lin, S.; Hu, H. Disentangled non-local neural networks. In: *Computer Vision–ECCV 2020. Lecture Notes in Computer Science, Vol. 12360*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 191–207, 2020.

[5] Hu, H.; Gu, J. Y.; Zhang, Z.; Dai, J. F.; Wei, Y. C. Relation networks for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3588–3597, 2018.

[6] Wang, X. L.; Girshick, R.; Gupta, A.; He, K. M. Non-local neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7794–7803, 2018.

[7] Hu, H.; Zhang, Z.; Xie, Z. D.; Lin, S. Local relation networks for image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 3463–3472, 2019.

[8] Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. OCNet: Object context network for scene parsing. *arXiv preprint* arXiv:1809.00916, 2018.

[9] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations, 2021.

[10] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the

Association for Computational Linguistics: Human Language Technologies, 4171–4186, 2019.

[11] Chen, M.; Radford, A.; Child, R.; Wu, J.; Jun, H.; Luan, D.; Sutskever, I. Generative pretraining from pixels. In: Proceedings of the 37th International Conference on Machine Learning, 1691–1703, 2020.

[12] Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; Douze, M. LeViT: A vision transformer in ConvNet's clothing for faster inference. *arXiv preprint* arXiv:2104.01136, 2021.

[13] Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient transformers: A survey. *arXiv preprint* arXiv:2009.06732, 2020.

[14] Liang, J.; Hu, D.; He, R.; Feng, J. Distill and fine-tune: Effective adaptation from a black-box source model. *arXiv preprint* arXiv:2104.01539, 2021.

[15] Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Tay, F. E.; Feng, J.; Yan, S. Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet. *arXiv preprint* arXiv:2101.11986, 2021.

[16] Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *arXiv preprint* arXiv:2103.00112, 2021.

[17] Chu, X. X.; Tian, Z.; Zhang, B.; Wang, X. L.; Wei, X. L.; Xia, H. X.; Shen, C. Conditional positional encodings for vision transformers. *arXiv preprint* arXiv:2102.10882, 2021.

[18] D'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; Sagun, L. ConViT: Improving vision transformers with soft convolutional inductive biases. In: Proceedings of the 38th International Conference on Machine Learning, 2286–2296, 2021.

[19] Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Hou, Q.; Feng, J. DeepViT: Towards deeper vision transformer. *arXiv preprint* arXiv:2103.11886, 2021.

[20] Liu, Z.; Lin, Y. T.; Cao, Y.; Hu, H.; Guo, B. N. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint* arXiv:2103.14030, 2021.

[21] Heo, B.; Yun, S.; Han, D.; Chun, S.; Oh, S. J. Rethinking spatial dimensions of vision transformers. *arXiv preprint* arXiv:2103.16302, 2021.

[22] Li, Y. W.; Zhang, K.; Cao, J. Z.; Timofte, R.; Gool, L. V. LocalViT: Bringing locality to vision transformers. *arXiv preprint* arXiv:2104.05707, 2021.

[23] Chefer, H.; Gur, S.; Wolf, L. Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 782–791, 2021.

[24] Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In: *Computer Vision–ECCV 2020. Lecture Notes in Computer Science, Vol. 12346*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 213–229, 2020.

[25] Zhu, X. Z.; Su, W. J.; Lu, L. W.; Li, B.; Dai, J. F. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv preprint* arXiv:2010.04159, 2020.

[26] Dai, Z. G.; Cai, B. L.; Lin, Y. G.; Chen, J. Y. UP-DETR: Unsupervised pre-training for object detection with transformers. *arXiv preprint* arXiv:2011.09094, 2020.

[27] Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao. L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint* arXiv:2102.12122, 2021.

[28] Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; Xia, H. End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8741–8750, 2021.

[29] Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint* arXiv:2105.15203, 2021.

[30] Kumar, M.; Weissenborn, D.; Kalchbrenner, N. Colorization transformer. In: Proceedings of the 9th International Conference on Learning Representations, 2021.

[31] Liu, B. C.; Song, K. P.; Zhu, Y. Z.; de Melo, G.; Elgammal, A. TIME: Text and image mutual-translation adversarial networks. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2082–2090, 2021.

[32] Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. *arXiv preprint* arXiv:2102.12092, 2021.

[33] Yang, F. Z.; Yang, H.; Fu, J. L.; Lu, H. T.; Guo, B. N. Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5790–5799, 2020.

[34] Jiang, Y. F.; Chang, S. Y.; Wang, Z. Y. TransGAN: Two transformers can make one strong GAN. *arXiv preprint* arXiv:2102.07074, 2021.

[35] Hudson, D. A.; Zitnick, C. L. Generative adversarial transformers. *arXiv preprint* arXiv:2103.01209, 2021.

[36] Van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6309–6318, 2017.

[37] Wang, Z.; Cun, X.; Bao, J.; Liu, J. Uformer: A general U-shaped transformer for image restoration. *arXiv preprint* arXiv:2106.03106, 2021.

[38] Deng, Y. Y.; Tang, F.; Pan, X. J.; Dong, W. M.; Xu, C. S. StyTr2: Unbiased image style transfer with transformers. *arXiv preprint* arXiv:2105.14576, 2021.

[39] Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R. R.; Hu, S.-M. PCT: Point cloud transformer. *Computational Visual Media* Vol. 7, No. 2, 187–199, 2021.

[40] Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd Conference on Neural Information Processing Systems, 13–23, 2019.

[41] Chen, Y.-C.; Li, L. J.; Yu, L. C.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. UNITER: UNiversal image-TExt representation learning. In: *Computer Vision–ECCV 2020. Lecture Notes in Computer Science, Vol. 12375*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 104–120, 2020.

[42] Li, C. L.; Yan, M.; Xu, H. Y.; Luo, F. L.; Huang, S. F. SemVLP: Vision-language pre-training by aligning semantics at multiple levels. *arXiv preprint* arXiv:2103.07829, 2021.

[43] Zhang, R.; Isola, P.; Efros, A. A. Colorful image colorization. In: *Computer Vision–ECCV 2016. Lecture Notes in Computer Science, Vol. 9907*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 649–666, 2016.

[44] Zhang, R.; Zhu, J.-Y.; Isola, P.; Geng, X. Y.; Lin, A. S.; Yu, T. H.; Efros, A. A. Real-time user-guided image colorization with learned deep priors. *arXiv preprint* arXiv:1705.02999, 2017.

[45] Su, J.-W.; Chu, H.-K.; Huang, J.-B. Instance-aware image colorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7965–7974, 2020.

[46] Pang, L.; Lan, Y.; Guo, J.; Xu, J.; Wan, S.; Cheng, X. Text matching as image recognition. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2793–2799, 2016.

[47] Dong, C.; Loy, C. C.; He, K. M.; Tang, X. O. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 38, No. 2, 295–307, 2016.

[48] Zhang, Y. L.; Tian, Y. P.; Kong, Y.; Zhong, B. N.; Fu, Y. Residual dense network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2472–2481, 2018.

[49] Haris, M.; Shakhnarovich, G.; Ukita, N. Deep back-projection networks for super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1664–1673, 2018.

[50] Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint* arXiv:1606.03657, 2016.

[51] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. *arXiv preprint* arXiv:1606.03498, 2016.

[52] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *arXiv preprint* arXiv:1706.08500, 2017.

[53] Karras, T.; Laine, S.; Aila, T. M. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4396–4405, 2019.

[54] Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein GANs. *arXiv preprint* arXiv:1704.00028, 2017.

[55] Bebis, G.; Georgiopoulos, M. Feed-forward neural networks. *IEEE Potentials* Vol. 13, No. 4, 27–31, 1994.

[56] Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer normalization. *arXiv preprint* arXiv:1607.06450, 2016.

[57] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010, 2017.

[58] Hendrycks, D.; Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint* arXiv:1606.08415, 2016.

[59] Kitaev, N.; Kaiser, L.; Levskaya, A. Reformer: The efficient transformer. In: Proceedings of the International Conference on Learning Representations, 2020.

[60] Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L. et al. Rethinking attention with performers. In: Proceedings of the International Conference on Learning Representations, 2021.

[61] Wang, S.; Li, B.; Khabsa, M.; Fang, H.; Ma, H. Linformer: Self-attention with linear complexity. *arXiv preprint* arXiv:2006.04768, 2020.

[62] Abnar, S.; Zuidema, W. Quantifying attention flow in transformers. *arXiv preprint* arXiv:2005.00928, 2020.

[63] Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5797–5808, 2019.

[64] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S. A.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* Vol. 115, No. 3, 211–252, 2015.

[65] Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning, 10347–10357, 2021.

[66] Han, Y. Z.; Huang, G.; Song, S. J.; Yang, L.; Wang, Y. L. Dynamic neural networks: A survey. *arXiv preprint* arXiv:2102.04906, 2021.

[67] Xu, W.; Xu, Y.; Chang, T.; Tu, Z. Co-scale conv-attentional image transformers. *arXiv preprint* arXiv:2104.06399, 2021.

[68] Dong, X. Y.; Bao, J. M.; Chen, D. D.; Zhang, W. M.; Yu, N. H.; Yuan, L.; Chen, D.; Guo, B. CSWin transformer: A general vision transformer backbone with cross-shaped windows. *arXiv preprint* arXiv:2107.00652, 2021.

[69] Huang, Z. L.; Wang, X. G.; Huang, L. C.; Huang, C.; Wei, Y. C.; Liu, W. CCNet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 603–612, 2019.

[70] Hou, Q. B.; Zhang, L.; Cheng, M. M.; Feng, J. S. Strip pooling: Rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4002–4011, 2020.

[71] Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; Jegou, H. Going deeper with image transformers. *arXiv preprint* arXiv:2103.17239, 2021.

[72] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, 618–626, 2017.

[73] Binder, A.; Montavon, G.; Lapuschkin, S.; Müller, K.-R.; Samek, W. Layer-wise relevance propagation for neural networks with local renormalization layers. In: *Artificial Neural Networks and Machine Learning–ICANN 2016. Lecture Notes in Computer Science, Vol. 9887*. Villa, A.; Masulli, P.; Pons Rivero, A. Eds. Springer Cham, 63–71, 2016.

[74] Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H. et al. Rethinking semantic segmentation from a sequence-tosequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6881–6890, 2021.

[75] Duke, B.; Ahmed, A.; Wolf, C.; Aarabi, P.; Taylor, G. W. SSTVOS: Sparse spatiotemporal transformers for video object segmentation. *arXiv preprint* arXiv:2101.08833, 2021.

[76] Chen, J. N.; Lu, Y. Y.; Yu, Q. H.; Luo, X. D.; Zhou, Y. Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint* arXiv:2102.04306, 2021.

[77] Ye, L. W.; Rochan, M.; Liu, Z.; Wang, Y. Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10494–10503, 2019.

[78] Wang, H.; Zhu, Y.; Adam, H.; Yuille, A.; Chen, L.-C. Max-deeplab: End-to-end panoptic segmentation with mask transformers. *arXiv preprint* arXiv:2012.00759, 2020.

[79] Durner, M.; Boerdijk, W.; Sundermeyer, M.; Friedl, W.; Marton, Z.-C.; Triebel, R. Unknown object segmentation from stereo images. *arXiv preprint* arXiv:2103.06796, 2021.

[80] Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 1, 172–186, 2021.

[81] Simon, T.; Joo, H.; Matthews, I.; Sheikh, Y. Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4645–4653, 2017.

[82] Cao, Z.; Simon, T.; Wei, S.-E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1302–1310, 2017.

[83] Fang, H.-S.; Xie, S. Q.; Tai, Y.-W.; Lu, C. W. RMPE: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, 2353–2362, 2017.

[84] Zhang, F.; Zhu, X. T.; Dai, H. B.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7091–7100, 2020.

[85] Sun, K.; Xiao, B.; Liu, D.; Wang, J. D. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5686–5696, 2019.

[86] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint* arXiv:1506.01497, 2015.

[87] Cai, Z. W.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 5, 1483–1498, 2021.

[88] Lin, T. Y.; Goyal, P.; Girshick, R.; He, K. M.; Dollár, P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2999–3007, 2017.

[89] Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv preprint* arXiv:1904.07850, 2019.

[90] Tian, Z.; Shen, C. H.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9626–9635, 2019.

[91] Stewart, R.; Andriluka, M.; Ng, A. Y. End-to-end people detection in crowded scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2325–2333, 2016.

[92] Hosang, J.; Benenson, R.; Schiele, B. Learning non-maximum suppression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6469–6477, 2017.

[93] Rezatofighi, S. H.; Kaskman, R.; Motlagh, F. T.; Shi, Q. F.; Cremers, D.; Leal-Taixé, L.; Reid, I. Deep perm-set net: Learn to predict sets with unknown permutation and cardinality using deep neural networks. *arXiv preprint* arXiv:1805.00613, 2018.

[94] Pan, X. J.; Tang, F.; Dong, W. M.; Gu, Y.; Song, Z. C.; Meng, Y. P.; Xu, P.; Deussen, O.; Xu, C. Self-supervised feature augmentation for large image object detection. *IEEE Transactions on Image Processing* Vol. 29, 6745–6758, 2020.

[95] Pan, X.; Gao, Y.; Lin, Z.; Tang, F.; Dong, W.; Yuan, H.; Huang, F.; Xu, C. Unveiling the potential of structure preserving for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11642–11651, 2021.

[96] Pan, X. J.; Ren, Y. Q.; Sheng, K. K.; Dong, W. M.; Yuan, H. L.; Guo, X. W.; Ma, C.; Xu, C. Dynamic refinement network for oriented and densely packed object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11204–11213, 2020.

[97] Chu, X. X.; Tian, Z.; Wang, Y. Q.; Zhang, B.; Shen, C. H. Twins: Revisiting spatial attention design in vision transformers. *arXiv preprint* arXiv:2104.13840, 2021.

[98] Beal, J.; Kim, E.; Tzeng, E.; Park, D. H.; Kislyuk, D. Toward transformer-based object detection. *arXiv preprint* arXiv:2012.09958, 2020.

[99] Dai, J. F.; Qi, H. Z.; Xiong, Y. W.; Li, Y.; Zhang, G. D.; Hu, H.; Wei, Y. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, 764–773, 2017.

[100] He, K. M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2980–2988, 2017.

[101] Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12299–12310, 2021.

[102] Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. *arXiv preprint* arXiv:2012.09841, 2020.

[103] Kaiser, L.; Bengio, S. Can active memory replace attention? In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 3781–3789, 2016.

[104] Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 4, 652–663, 2016.

[105] Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3156–3164, 2015.

[106] Rolfe, J. T. Discrete variational autoencoders. *arXiv preprint* arXiv:1609.02200, 2016.

[107] Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv preprint* arXiv:1406.2661, 2014.

[108] Ho, J.; Kalchbrenner, N.; Weissenborn, D.; Salimans, T. Axial attention in multidimensional transformers. *arXiv preprint* arXiv:1912.12180, 2019.

[109] Antol, S.; Agrawal, A.; Lu, J. S.; Mitchell, M.; Batra, D.; Zitnick, C. L.; Parikh, D. VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision, 2425–2433, 2015.

[110] Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6325–6334, 2017.

[111] Chen, X. L.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; Zitnick, C. L. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint* arXiv:1504.00325, 2015.

[112] Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* Vol. 2, 67–78, 2014.

[113] Gan, Z.; Chen, Y.-C.; Li, L.; Zhu, C.; Cheng, Y.; Liu, J. Large-scale adversarial training for vision-and-language representation learning. In: *Advances in Neural Information Processing Systems, Vol. 33*. Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; Lin, H. Eds. Curran Associates, Inc., 6616–6628, 2020.

[114] Lin, J. Y.; Yang, A.; Zhang, Y. C.; Liu, J.; Yang, H. X. InterBERT: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint* arXiv:2003.13198, 2020.

[115] Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. VL-BERT: Pre-training of generic visual-linguistic representations. In: Proceedings of the International Conference on Learning Representations, 2020.

[116] Zhou, L. W.; Palangi, H.; Zhang, L.; Hu, H. D.; Corso, J.; Gao, J. F. Unified vision-language pre-training for image captioning and VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 7, 13041–13049, 2020.

[117] Girshick, R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 1440–1448, 2015.

[118] Li, W.; Gao, C.; Niu, G. C.; Xiao, X. Y.; Wang, H. F. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint* arXiv:2012.15409, 2020.

[119] Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C. J.; Chang, K. W. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint* arXiv:1908.03557, 2019.

[120] Alberti, C.; Ling, J.; Collins, M.; Reitter, D. Fusion of detected objects in text for visual question answering. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2131–2140, 2019.

[121] Li, X. J.; Yin, X.; Li, C. Y.; Zhang, P. C.; Hu, X. W.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F. et al. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In: *Computer Vision–ECCV 2020. Lecture Notes in Computer Science, Vol. 12375*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 121–137, 2020.

[122] Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; Wang, H. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[123] Li, Y.; Pan, Y.; Yao, T.; Chen, J.; Mei, T. Scheduled sampling in vision-language pretraining with decoupled encoder–decoder network. In: Proceedings of the AAAI Conference on Artificial Intelligence, 8518–8526, 2021.

[124] Tan, H.; Bansal, M. LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 5100–5111, 2019.

[125] Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2556–2565, 2018.

[126] Ordonez, V.; Kulkarni, G.; Berg, T. L. Im2Text: Describing images using 1 million captioned photographs. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, 1143–1151, 2011.

[127] Krishna, R.; Zhu, Y. K.; Groth, O.; Johnson, J.; Hata, K. J.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A. et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* Vol. 123, No. 1, 32–73, 2017.

[128] Hudson, D. A.; Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6693–6702, 2019.

[129] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D., Dollar, P.; Zitnick, C. L. Microsoft COCO: Common objects in context. In: *Computer Vision–ECCV 2014. Lecture Notes in Computer Science, Vol. 8693*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 740–755, 2014.

[130] Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A. et al. The open images dataset V4. *International Journal of Computer Vision* Vol. 128, No. 7, 1956–1981, 2020.

[131] Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. VinVL: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5579–5588, 2021.

[132] Hu, R.; Singh, A. UniT: Multimodal multitask learning with a unified transformer. *arXiv preprint* arXiv:2102.10772, 2021.

[133] Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H. J.; Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6418–6428, 2019.

[134] Xie, N.; Lai, F.; Doran, D.; Kadav, A. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint* arXiv:1901.06706, 2019.

[135] Zellers, R.; Bisk, Y.; Farhadi, A.; Choi, Y. From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6713–6724, 2019.

[136] Kazemzadeh, S.; Ordonez, V.; Matten, M.; Berg, T. ReferItGame: Referring to objects in photographs of natural scenes. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 787–798, 2014.

[137] Sheng, K. K.; Dong, W. M.; Ma, C. Y.; Mei, X.; Huang, F. Y.; Hu, B.-G. Attention-based multi-patch aggregation for image aesthetic assessment. In: Proceedings of the 26th ACM International Conference on Multimedia, 879–886, 2018.

[138] Sheng, K. K.; Dong, W. M.; Chai, M. L.; Wang, G. H.; Zhou, P.; Huang, F. Y.; Hu, B.-G.; Ji, R.; Ma, C. Revisiting image aesthetic assessment via self-supervised feature learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 4, 5709–5716, 2020.

[139] Sheng, K. K.; Dong, W. M.; Huang, H. B.; Chai, M. L.; Zhang, Y.; Ma, C. Y.; Hu, B.-G. Learning to assess visual aesthetics of food images. *Computational Visual Media* Vol. 7, No. 1, 139–152, 2021.

[140] Zhang, S. F.; Wang, X. B.; Liu, A.; Zhao, C. X.; Wan, J.; Escalera, S.; Shi, H.; Wang, Z.; Li, S. Z. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 919–928, 2019.

[141] Chen, Z.; Yao, T.; Sheng, K.; Ding, S.; Tai, Y.; Li, J.; Huang, F.; Jin, X. Generalizable representation learning for mixture domain face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence, 1132–1139, 2021.

[142] Zhao, H.; Jiang, L.; Jia, J.; Torr, P.; Koltun, V. Point transformer. *arXiv preprint* arXiv:2012.09164, 2020.

[143] Zoph, B.; Le, Q. V. Neural architecture search with reinforcement learning. In: Proceedings of the International Conference on Learning Representations, 2017.

[144] Zoph, B.; Vasudevan, V.; Shlens, J.; Le, Q. V. Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, 8697–8710, 2018.

[145] Real, E.; Aggarwal, A.; Huang, Y. P.; Le, Q. V. Regularized evolution for image classifier architecture search. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 4780–4789, 2019.

[146] Wang, H. R.; Wu, Z. H.; Liu, Z. J.; Cai, H.; Zhu, L. G.; Gan, C.; Han, S. HAT: Hardware-aware transformers for efficient natural language processing. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 7675–7688, 2020.

[147] So, D.; Le, Q.; Liang, C. The evolved transformer. In: Proceedings of the 36th International Conference on Machine Learning, 5877–5886, 2019.

[148] Li, C. L.; Tang, T.; Wang, G. R.; Peng, J. F.; Chang, X. J. BossNAS: Exploring hybrid CNN-transformers with Block-wisely Self-supervised neural architecture search. *arXiv preprint* arXiv:2103.12424, 2021.

[149] Schulz, K.; Sixt, L.; Tombari, F.; Landgraf, T. Restricting the flow: Information bottlenecks for attribution. In: Proceedings of the International Conference on Learning Representations, 2019.

[150] Jiang, Z.; Tang, R.; Xin, J.; Lin, J. Inserting information bottleneck for attribution in transformers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: Findings, 3850–3857, 2020.

Yifan Xu is currently a postgraduate of the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation, Chinese Academy of Sciences. He received a his B.Eng. degree from Beijing Institute of Technology in 2015. His research interests include transfer learning, machine learning, and computational visual media.

Huapeng Wei is a postgraduate of the School of Artificial Intelligence, Jilin University. He received his B.Sc. degree from Jilin University in 2020. His research interests include computational visual media and image processing.

Minxuan Lin received his B.Sc. degree in computer science and technology from the Ocean University of China in 2018. He is currently a postgraduate of NLPR. His research interests include computational visual media and machine learning.

Yingying Deng received her B.Sc. degree in automation from the University of Science and Technology, Beijing in 2017. She is currently working towards her Ph.D. degree in NLPR. Her research interests include computational visual media and machine learning.

Kekai Sheng received his Ph.D. degree from NLPR in 2019. He received his B.Eng. degree in telecommunication engineering from the University of Science and Technology, Beijing in 2014. He is currently a research engineer at Youtu Lab, Tencent Inc. His research interests include domain adaptation, neural architecture search, and AutoML.

Mengdan Zhang received her Ph.D. degree from NLPR in 2018. She received her B.Eng. degree in automation from Xi'an Jiao Tong University in 2013. She is currently a research engineer at Youtu Lab, Tencent Inc. Her research interests include computer vision and machine learning.

Fan Tang is an assistant professor in the School of Artificial Intelligence, Jilin University. He received his B.Sc. degree in computer science from North China Electric Power University in 2013 and his Ph.D. degree from NLPR in 2019. His research interests include computer graphics, computer vision, and machine learning.

Weiming Dong is a professor in NLPR. He received his B.Eng. and M.S. degrees in computer science in 2001 and 2004 from Tsinghua University. He received his Ph.D. degree in information technology from the University of Lorraine, France, in 2007. His research interests include visual media synthesis and evaluation. Weiming Dong is a member of the ACM and IEEE.

Feiyue Huang is the director of the Youtu Lab, Tencent Inc. He received his B.Sc. and Ph.D. degrees in computer science in 2001 and 2008 respectively, both from Tsinghua University, China. His research interests include image understanding and face recognition.

Changsheng Xu is a professor in NLPR. His research interests include multimedia content analysis, indexing and retrieval, pattern recognition, and computer vision. Prof. Xu has served as associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 prestigious IEEE and ACM multimedia journals, conferences, and workshops. Currently he is the editor-in-chief of *Multimedia Systems*. Changsheng Xu is an IEEE Fellow, IAPR Fellow, and ACM Distinguished Scientist.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit `http://creativecommons.org/licenses/by/4.0/`.

Other papers from this open access journal are available free of charge from http://www.springer.com/journal/41095. To submit a manuscript, please go to https://www.editorialmanager.com/cvmj.