

Transformers in Vision: A Survey

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir,
Fahad Shahbaz Khan, and Mubarak Shah

Abstract—Astounding results from Transformer models on natural language tasks have intrigued the vision community to study their application to computer vision problems. Among their salient benefits, Transformers enable modeling long dependencies between input sequence elements and support parallel processing of sequence as compared to recurrent networks e.g., Long short-term memory (LSTM). Different from convolutional networks, Transformers require minimal inductive biases for their design and are naturally suited as set-functions. Furthermore, the straightforward design of Transformers allows processing multiple modalities (e.g., images, videos, text and speech) using similar processing blocks and demonstrates excellent scalability to very large capacity networks and huge datasets. These strengths have led to exciting progress on a number of vision tasks using Transformer networks. This survey aims to provide a comprehensive overview of the Transformer models in the computer vision discipline. We start with an introduction to fundamental concepts behind the success of Transformers i.e., self-attention, large-scale pre-training, and bidirectional feature encoding. We then cover extensive applications of transformers in vision including popular recognition tasks (e.g., image classification, object detection, action recognition, and segmentation), generative modeling, multi-modal tasks (e.g., visual-question answering, visual reasoning, and visual grounding), video processing (e.g., activity recognition, video forecasting), low-level vision (e.g., image super-resolution, image enhancement, and colorization) and 3D analysis (e.g., point cloud classification and segmentation). We compare the respective advantages and limitations of popular techniques both in terms of architectural design and their experimental value. Finally, we provide an analysis on open research directions and possible future works. We hope this effort will ignite further interest in the community to solve current challenges towards the application of transformer models in computer vision.

Index Terms—Self-attention, transformers, bidirectional encoders, deep neural networks, convolutional networks, self-supervision.

1 INTRODUCTION

TRANSFORMER models [1] have recently demonstrated exemplary performance on a broad range of language tasks e.g., text classification, machine translation [2] and question answering. Among these models, the most popular ones include BERT (Bidirectional Encoder Representations from Transformers) [3], GPT (Generative Pre-trained Transformer) v1-3 [4]–[6], RoBERTa (Robustly Optimized BERT Pre-training) [7] and T5 (Text-to-Text Transfer Transformer) [8]. The profound impact of Transformer models has become more clear with their scalability to very large capacity models [9], [10]. For example, the BERT-large [3] model with 340 million parameters was significantly outperformed by the GPT-3 [6] model with 175 billion parameters while the latest mixture-of-experts Switch transformer [10] scales up to a whopping 1.6 trillion parameters!

The breakthroughs from Transformer networks in Natural Language Processing (NLP) domain has sparked great interest in the computer vision community to adapt these models for vision and multi-modal learning tasks (Fig. 1).

However, visual data follows a typical structure (e.g., spatial and temporal coherence), thus demanding novel network designs and training schemes. As a result, Transformer models and their variants have been successfully used for image recognition [11], [12], object detection [13], [14], segmentation [15], image super-resolution [16], video understanding [17], [18], image generation [19], text-image synthesis [20] and visual question answering [21], [22], among several other use cases [23]–[26]. This survey aims to cover such recent and exciting efforts in the computer vision domain, providing a comprehensive reference to interested readers.

Transformer architectures are based on a self-attention mechanism that learns the relationships between elements of a sequence. As opposed to recurrent networks that process sequence elements recursively and can only attend to short-term context, Transformers can attend to complete sequences thereby learning long-range relationships. Although attention models have been extensively used in both feed-forward and recurrent networks [27], [28], Transformers are based solely on the attention mechanism and have a unique implementation (i.e., multi-head attention) optimized for parallelization. An important feature of these models is their scalability to high-complexity models and large-scale datasets e.g., in comparison to some of the other alternatives such as hard attention [29] which is stochastic in nature and requires Monte Carlo sampling for sampling attention locations. Since Transformers assume minimal prior knowledge about the structure of the problem as compared to their convolutional and recurrent counterparts [30]–[32], they are typically pre-trained using pretext tasks on large-scale (unlabelled) datasets [1], [3]. Such a pre-training avoids costly manual annotations, thereby encoding highly expres-

- S. Khan, M. Naseer and F. S. Khan are with the MBZ University of Artificial Intelligence, Abu Dhabi, UAE.
E-mail: firstname.lastname@mbzuaei.ac.ae
- M. Hayat is with the Faculty of IT, Monash University, Clayton VIC 3800, Australia.
- S. W. Zamir is with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE.
- S. Khan and M. Naseer are also with the CECS, Australian National University, Canberra ACT 0200, Australia.
- F. S. Khan is also with the Computer Vision Laboratory, Linköping University, Sweden.
- M. Shah is with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816, United States.

Manuscript received March, 2021.

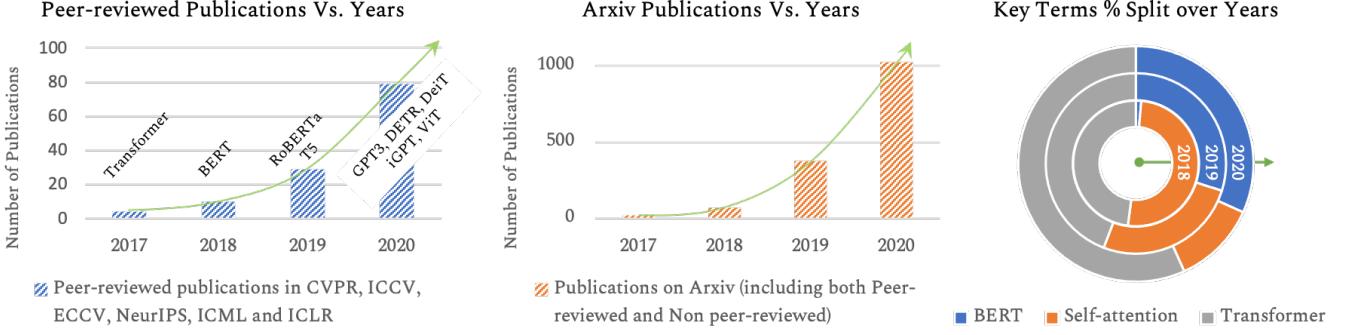


Fig. 1: Statistics on the number of times keywords such as BERT, Self-Attention, and Transformers appear in the titles of Peer-reviewed and arXiv papers over the past few years (in Computer Vision and Machine Learning). The plots show consistent growth in recent literature. This survey covers recent progress on Transformers in the computer vision domain.

sive and generalizable representations that model rich relationships between the entities present in a given dataset. The learned representations are then fine-tuned on the downstream tasks in a supervised manner to obtain favorable results.

This paper provides a holistic overview of the transformer models developed for computer vision applications. We develop a taxonomy of the network design space and highlight the major strengths and shortcomings of the existing methods. Other literature reviews mainly focus on the NLP domain [33], [34] or cover generic attention-based approaches [27], [33]. By focusing on the newly emerging area of visual transformers, we comprehensively organize the recent approaches according to the intrinsic features of self-attention and the investigated task. We first provide an introduction to the salient concepts underlying Transformer networks and then elaborate on the specifics of recent vision transformers. Where ever possible, we draw parallels between the Transformers used in the NLP domain [1] and the ones developed for vision problems to flash major novelties and interesting domain-specific insights. Recent approaches show that convolution operations can be fully replaced with attention-based transformer modules and have also been used jointly in a single design to encourage symbiosis between the two complementary set of operations. This survey finally details open research questions with an outlook towards the possible future work.

2 FOUNDATIONS

There exist two key ideas that have contributed towards the development of conventional transformer models. (a) The first one is *self-attention*, which allows capturing ‘long-term’ dependencies between sequence elements as compared to conventional recurrent models that find it challenging to encode such relationships. (b) The second key idea is that of *pre-training*¹ on a large (un)labelled corpus in a (self)supervised manner, and subsequently fine-tuning to the target task with a small labeled dataset [3], [7], [38]. Below, we provide a brief tutorial on these two ideas (Sec. 2.2 and 2.1), along with a summary of seminal Transformer

1. Several recent Vision Transformers demonstrate that the model can be learned end-to-end on ImageNet-1K without any dedicated pre-training phase [35]–[37]. However, the performance generally remains lower than the pre-trained counter-parts.

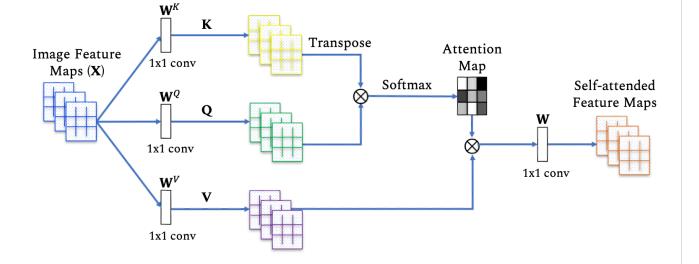


Fig. 2: An example self-attention block used in the vision domain [39]. Given the input sequence of image features, the triplet of (key, query, value) is calculated followed by attention calculation and applying it to reweight the values. A single head is shown here and an output projection (W) is finally applied to obtain output features with the same dimension as the input. Figure adapted from [39].

networks (Sec. 2.3 and 2.4) where these ideas have been applied. This background will help us better understand the forthcoming Transformer based models used in the computer vision domain (Sec. 3).

2.1 Self-Attention in Transformers

Given a sequence of items, self-attention estimates the relevance of one item to other items (e.g., which words are likely to come together in a sentence). The self-attention mechanism is an integral component of Transformers, which explicitly models the interactions between all entities of a sequence for structured prediction tasks. Basically, a self-attention layer updates each component of a sequence by aggregating global information from the complete input sequence. Let's denote a sequence of n entities $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ by $\mathbf{X} \in \mathbb{R}^{n \times d}$, where d is the embedding dimension to represent each entity. The goal of self-attention is to capture the interaction amongst all n entities by encoding each entity in terms of the global contextual information. This is done by defining three learnable weight matrices to transform Queries ($\mathbf{W}^Q \in \mathbb{R}^{d \times d_q}$), Keys ($\mathbf{W}^K \in \mathbb{R}^{d \times d_k}$) and Values ($\mathbf{W}^V \in \mathbb{R}^{d \times d_v}$), where $d_q = d_k$. The input sequence \mathbf{X} is first projected onto these weight matrices to get $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$ and $\mathbf{V} = \mathbf{X}\mathbf{W}^V$. The output $\mathbf{Z} \in \mathbb{R}^{n \times d_v}$ of the self attention layer is,

$$\mathbf{Z} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_q}} \right) \mathbf{V}.$$

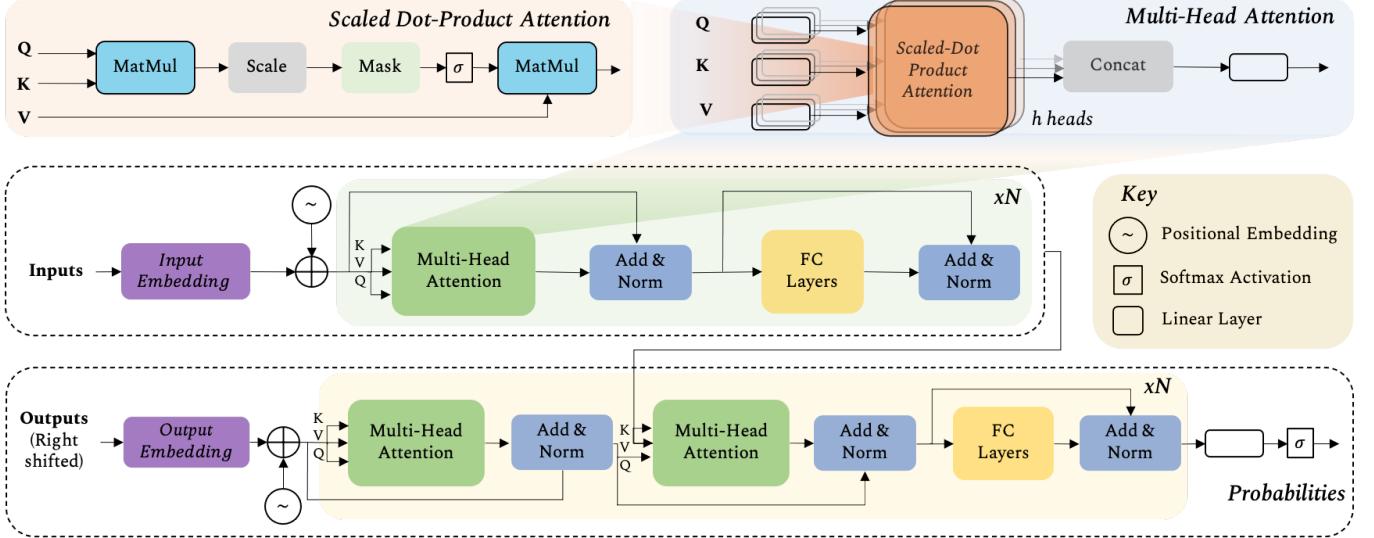


Fig. 3: Architecture of the Transformer Model [1]. The model was first developed for the language translation task where an input sequence in one language is required to be converted to the output sequence in another language. The Transformer encoder (middle row) operates on the input language sequence and converts it to an embedding before passing it on to the encoder blocks. The Transformer decoder (bottom row) operates on the previously generated outputs in the translated language and the encoded input sequence from the middle branch to output the next word in the output sequence. The sequence of previous outputs (used as input to the decoder) is obtained by shifting the output sentence to the right by one position and appending start-of-sentence token at the beginning. This shifting avoids the model to learn to simply copy the decoder input to the output. The ground-truth to train the model is simply the output language sequence (without any right shift) appended with an end-of-sentence token. The blocks consisting of multi-head attention (top row) and feed-forward layers are repeated N times in both the encoder and decoder.

For a given entity in the sequence, the self-attention basically computes the dot-product of the query with all keys, which is then normalized using softmax operator to get the attention scores. Each entity then becomes the weighted sum of all entities in the sequence, where weights are given by the attention scores (Fig. 2 and Fig. 3, top row-left block).

Masked Self-Attention: The standard self-attention layer attends to all entities. For the Transformer model [1] which is trained to predict the next entity of the sequence, the self-attention blocks used in the decoder are masked to prevent attending to the subsequent future entities. This is simply done by an element-wise multiplication operation with a mask $M \in \mathbb{R}^{n \times n}$, where M is an upper-triangular matrix. The masked self-attention is defined by,

$$\text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_q}} \circ M \right),$$

where \circ denotes Hadamard product. Basically, while predicting an entity in the sequence, the attention scores of the future entities are set to zero in masked self-attention.

Multi-Head Attention: In order to encapsulate multiple complex relationships amongst different elements in the sequence, the multi-head attention comprises multiple self-attention blocks ($h = 8$ in the original Transformer model [1]). Each block has its own set of learnable weight matrices $\{\mathbf{W}^{Q_i}, \mathbf{W}^{K_i}, \mathbf{W}^{V_i}\}$, where $i = 0 \dots (h-1)$. For an input \mathbf{X} , the output of the h self-attention blocks in multi-head attention is then concatenated into a single matrix $[\mathbf{Z}_0, \mathbf{Z}_1, \dots, \mathbf{Z}_{h-1}] \in \mathbb{R}^{n \times h \cdot d_v}$ and projected onto a weight matrix $\mathbf{W} \in \mathbb{R}^{h \cdot d_v \times d}$ (Fig. 3, top row).

The main difference of self-attention with convolution operation is that the filters are dynamically calculated in

stead of static filters (that stay the same for any input) as in the case of convolution. Further, self-attention is invariant to permutations and changes in the number of input points. As a result, it can easily operate on irregular inputs as opposed to standard convolution that requires grid structure. Furthermore, it has been shown in the literature how self-attention (with positional encodings) is theoretically a more flexible operation which can model the behaviour of convolutional models towards encoding local features [40]. Cordonnier *et al.* [41] further studied the relationships between self-attention and convolution operations. Their empirical results confirm that multi-head self-attention (with sufficient parameters) is a more generic operation which can model the expressiveness of convolution as a special case. In fact, self-attention provides the capability to learn the global as well as local features, and provide expressivity to adaptively learn kernel weights as well as the receptive field (similar to deformable convolutions [42]).

2.2 (Self) Supervised Pre-training

Self-attention based Transformer models generally operate in a two-stage training mechanism. First, pre-training is performed on a large-scale dataset (and sometimes a combination of several available datasets [22], [43]) in either a supervised [11] or a self-supervised manner [3], [44], [45]. Later, the pre-trained weights are adapted to the downstream tasks using small-mid scale datasets. Examples of downstream tasks include image classification [46], object detection [13], zero-shot classification [20], question-answering [10] and action recognition [18]. The effectiveness of pre-training for large-scale Transformers has been advocated in both the language and vision domains. For

example, Vision Transformer model (ViT-L) [11] experiences an absolute 13% drop in accuracy on ImageNet test set when trained only on ImageNet train set as compared to the case when pretrained on JFT dataset [47] with 300 million images.

Since acquiring manual labels at a massive scale is cumbersome, self-supervised learning has been very effectively used in the pre-training stage. The self-supervision based pre-training stage training has played a crucial role in unleashing the scalability and generalization of Transformer networks, enabling training even above a *trillion* parameter networks (e.g., the latest Switch Transformer [10] from Google). An extensive survey on SSL can be found in [48], [49]. As nicely summarized by Y. LeCun [50], the basic idea of SSL is to *fill in the blanks*, i.e., try to predict the occluded data in images, future or past frames in temporal video sequences or predict a pretext task e.g., the amount of rotation applied to inputs, the permutation applied to image patches or the color of a gray-scale image. Another effective way to impose self-supervised constraints is via contrastive learning. In this case, nuisance transformations are used to create two types of modified versions of the same image i.e., without changing the underlying class semantics (e.g., image stylizing, cropping) and with semantic changes (e.g., replacing an object with another in the same scene, or changing the class with minor adversarial changes to the image). Subsequently, the model is trained to be invariant to the nuisance transformations and emphasize on modeling minor changes that can alter semantic labels.

Self-supervised learning provides a promising learning paradigm since it enables learning from a vast amount of readily available non-annotated data. In the SSL based pre-training stage, a model is trained to learn a meaningful representation of the underlying data by solving a pretext task. The pseudo-labels for the pretext task are automatically generated (without requiring any expensive manual annotations) based on data attributes and task definition. Therefore, the pretext task definition is a critical choice in SSL. We can broadly categorize existing SSL methods based upon their pretext tasks into **(a)** *generative* approaches which synthesize images or videos (given conditional inputs), **(b)** *context-based* methods which exploit the relationships between image patches or video frames, and **(c)** *cross-modal* methods which leverage from multiple data modalities. Examples of *generative* approaches include conditional generation tasks such as masked image modeling [43] and image colorization [51], image super-resolution [52], image in-painting [53], and GANs based methods [54], [55]. The *context-based* pretext methods solve problems such as a jigsaw puzzle on image patches [56]–[58], masked object classification [22], predict geometric transformation such as rotation [46], [59], or verify temporal sequence of video frames [60]–[62]. Cross-modal pretext methods verify the correspondence of two input modalities e.g., text & image [63], audio & video [64], [65] or RGB & flow [66].

2.3 Transformer Model

The architecture of the Transformer model proposed in [1] is shown in Fig. 3. It has an encoder-decoder structure. The encoder (*middle row*) consists of six identical blocks (i.e.,

$N=6$ in Fig. 3), with each block having two sub-layers: a multi-head self-attention network, and a simple position-wise fully connected feed-forward network. Residual connections [67] alongside layer normalization [68] are employed after each block as in Fig. 3. Note that, different from regular convolutional networks where feature aggregation and feature transformation are simultaneously performed (e.g., with a convolution layer followed by a non-linearity), these two steps are decoupled in the Transformer model i.e., self-attention layer only performs aggregation while the feed-forward layer performs transformation. Similar to the encoder, the decoder (*bottom row*) in the Transformer model comprises six identical blocks. Each decoder block has three sub-layers, first two (multi-head self-attention, and feed-forward) are similar to the encoder, while the third sub-layer performs multi-head attention on the outputs of the corresponding encoder block, as shown in Fig. 3.

The original Transformer model in [1] was trained for the Machine Translation task. The input to the encoder is a sequence of words (sentence) in one language. **Positional encodings** are added to the input sequence to capture the relative position of each word in the sequence. Positional encodings have the same dimensions as the input $d = 512$, and can be learned or pre-defined e.g., by sine or cosine functions. Being an auto-regressive model, the decoder of the Transformer [1] uses previous predictions to output the next word in the sequence. The decoder, therefore, takes inputs from the encoder as well as the previous outputs to predict the next word of the sentence in the translated language. To facilitate residual connections the output dimensions of all layers are kept the same i.e., $d = 512$. The dimensions of query, key and value weight matrices in multi-head attention are set to $d_q = 64$, $d_k = 64$, $d_v = 64$.

2.4 Bidirectional Representations

The training strategy of the original Transformer model [1] could only attend to the context on the left of a given word in the sentence. This is limiting, since for most language tasks, contextual information from both left and right sides is important. Bidirectional Encoder Representations from Transformers (BERT) [3] proposed to jointly encode the right and left context of a word in a sentence, thus improving the learned feature representations for textual data in an self-supervised manner. To this end, BERT [3] introduced two pretext tasks to pre-train the Transformer model [1] in a self-supervised manner: *Masked Language Model* and *Next Sentence Prediction*. For adapting the pre-trained model for downstream tasks, a task-specific additional output module is appended to the pre-trained model, and the full model is fine-tuned end-to-end. Here, we briefly touch upon the pretext tasks. **(1) Masked Language Model (MLM)** - A fixed percentage (15%) of words in a sentence are randomly masked and the model is trained to predict these masked words using cross-entropy loss. In predicting the masked words, the model learns to incorporate the bidirectional context. **(2) Next Sentence Prediction (NSP)** - Given a pair of sentences, the model predicts a binary label i.e., whether the pair is valid from the original document or not. The training data for this can easily be generated from any monolingual text corpus. A pair of sentences A and B is

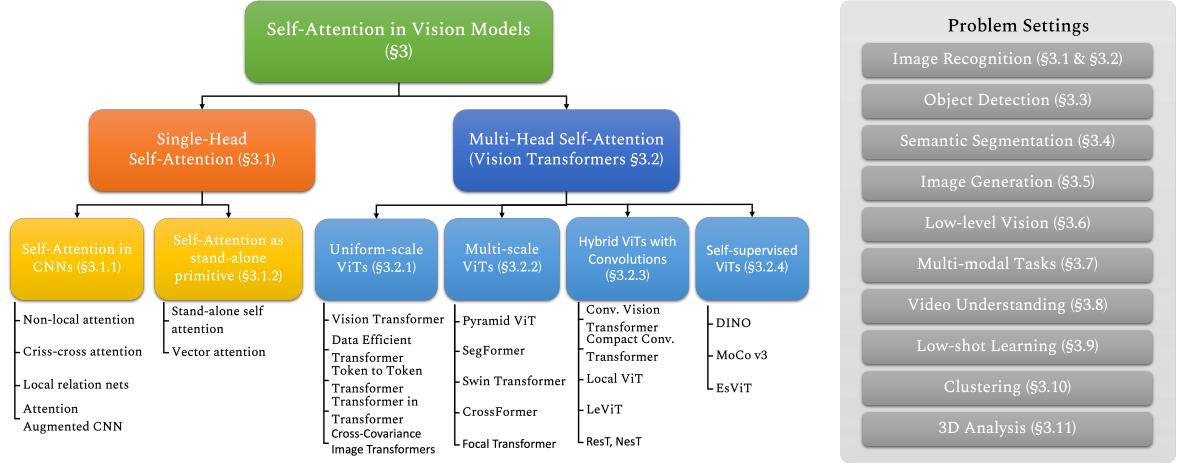


Fig. 4: A taxonomy of self-attention design space. Existing approaches based on self-attention explore single-head or multi-head (transformer) designs for vision tasks. We note that interesting efforts have been made to utilize knowledge from convolution based architectures to improve ViTs (e.g., multi-scale and hybrid designs). We categorize the upcoming sections of this survey according to the types of self-attention block (left tree diagram) as well as the prominent tasks in computer vision (right).

formed, such that B is the actual sentence (next to A) 50% of the time, and B is a random sentence for other 50% of the time. NSP enables the model to capture sentence-to-sentence relationships which are crucial in many language modeling tasks such as Question Answering and Natural Language Inference.

3 SELF-ATTENTION & TRANSFORMERS IN VISION

We broadly categorize vision models with self-attention into two categories: the models which use single-head self-attention (Sec. 3.1), and the models which employ multi-head self-attention based Transformer modules into their architectures (Sec. 3.2). Below, we first discuss the first category of single-head self-attention based frameworks, which generally apply global or local self-attention within CNN architectures, or utilize matrix factorization to enhance design efficiency and use vectorized attention models. We then discuss the Transformer-based vision architectures in Sec. 3.2.

3.1 Single-head Self-Attention

3.1.1 Self-Attention in CNNs

Inspired by non-local means operation [69] which was mainly designed for image denoising, Wang *et al.* [70] proposed a differentiable non-local operation for deep neural networks to capture long-range dependencies both in space and time in a feed-forward fashion. Given a feature map, their proposed operator [70] computes the response at a position as a weighted sum of the features at all positions in the feature map. This way, the non-local operation is able to capture interactions between any two positions in the feature map regardless of the distance between them. Videos classification is an example of a task where long-range interactions between pixels exist both in space and time. Equipped with the capability to model long-range interactions, [70] demonstrated the superiority of non-local deep neural networks for more accurate video classification on Kinetics dataset [71].

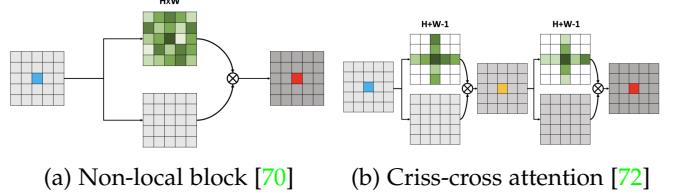


Fig. 5: Comparison of two different self-attention approaches: Non-local self-attention block [70] and Criss-cross self-attention module [72]. Figure is from [72].

Although the self-attention allows us to model full-image contextual information, it is both memory and compute intensive. As shown in Fig. 5(a), in order to encode global context for a given pixel location, non-local block [70] computes a *dense* attention map (in green). The non-local block [70] has a high complexity of $\mathcal{O}(N^2)$, where N denotes the number of input feature maps. To reduce this computational burden, Huang *et al.* [72] propose the criss-cross attention module that for each pixel position generates a *sparse* attention map only on the criss-cross path, as illustrated in Fig. 5(b). Further, by applying criss-cross attention recurrently, each pixel position can capture context from all other pixels. Compared to non-local block, the criss-cross uses $11 \times$ lesser GPU memory, and has a complexity of $\mathcal{O}(2\sqrt{N})$. State-of-the-art results are reported [72] for the semantic and instance segmentation tasks on several benchmark datasets including Cityscapes [73], ADE20K [74], COCO [75], LIP [76] and CamVid [77].

Another shortcoming of the convolutional operator comes from the fact that after training, it applies fixed weights regardless of any changes to the visual input. Hu *et al.* [78] proposed local relation networks to adaptively compose pixels in a local window. They introduced a new differentiable layer that adapts its weight aggregation based on the compositional relations (similarity) between pixels/features within a local window. Such adaptive weight aggregation introduces geometric priors into the network which are important for the recognition tasks [78]. Convolution is considered to be a top-down operator as it remains

fixed across positions while a non-local operation such as introduced in [69] is a bottom-up method as it aggregates input features over the full image. The local relation layer belongs to the category of bottom-up methods but it is restricted to a fixed window size *e.g.*, 7x7 neighborhood.

Bello *et al.* [79] explore the possibility of employing self-attention as an alternative to convolutional operators. They employ the relative position encoding [80] in two dimensions to develop a new self-attention mechanism that maintains translation equivariance, a desirable property for handling images. Although this self-attention provides competitive results as a stand-alone computational primitive, the best performance is obtained in combination with the convolutional operations. Authors show that attention augmentation leads to systematic performance gains in image classification and object detection for different architectures.

3.1.2 Self-Attention as Stand-alone Primitive

As discussed above, convolutional layers possess translation equivariance but can not scale with a large receptive field, therefore can not capture long-range interactions [81]. On the other hand, global attention [1] which attend to all spatial locations of the input can be computationally intensive and is preferred on down-sampled small images, image patches [11] or augmenting the convolutional features space [79]. Ramachandran *et al.* [81] proposed to replace convolutional layers in deep neural networks with a local self-attention layer which can be applied to small or large inputs without increasing the computational cost. At a basic level, the proposed self-attention layer [81] considers all pixel positions in a specific window size around a given pixel, compute queries, keys and value vectors for these pixels, and then aggregates the spatial information within this window. The value vectors are aggregated after projecting the softmax score of queries and keys. This process is repeated for all given pixels and the response is concatenated to produce the output pixel. ResNet models with local self-attention layer can solve ImageNet and COCO object detection with fewer parameters as compared to ResNet models based on convolutional layers [81].

Zhao *et al.* [82] note that a traditional convolution operator performs feature aggregation and transformation jointly (by applying a filter and then passing it through a non-linearity). In contrast, they propose to perform feature aggregation separately with self-attention followed by transformation using an element-wise perceptron layer. For feature aggregation, they propose two alternate strategies: (a) pairwise self-attention and (b) patch-wise self-attention. The pairwise self-attention is permutation and cardinality invariant operation, while the patch-wise self-attention does not have such invariance properties (similar to convolution). Both pairwise and patch-wise self-attentions are implemented as a *vector* attention [82] that learns weights for both the spatial and channel dimensions. This provides an alternate approach for attention that is conventionally performed using scalar weights (by taking a dot-product). The pairwise self-attention is a set operator that computes a *vector attention* keeping in view the relationships of a particular feature with its neighbors in a given local neighborhood. In contrast, patch-wise self-attention is a generalization of the convolution operator (not a set operator) and looks at

all the feature vectors in the local neighbourhood when deriving the attention vectors. Authors show that with considerably fewer parameters, self-attention networks (SAN) can beat ResNet baselines on the ImageNet dataset. They further show robustness against adversarial perturbations [83], [84] and generalization to unseen transformations [85]. This behaviour is due to the dynamic nature of attention that makes it difficult for the adversary to calculate useful fooling directions.

3.2 Multi-head Self-Attention (Transformers)

Unlike the approaches discussed in Sec. 3.1 which insert self-attention as a component in CNN inspired architectures, Vision Transformer (ViTs) [11] adapts the architecture of [1] (see Fig. 3), which cascades multiple Transformer layers. ViTs have gained significant research attention, and a number of recent approaches have been proposed which build upon ViTs. Below, we discuss these methods by categorizing them into: uniform scale ViTs having single-scale features through all layers (Sec. 3.2.1), multi-scale ViTs that learn hierarchical features which are more suitable for dense prediction tasks (Sec. 3.2.2), and hybrid designs having convolution operations within ViTs (Sec. 3.2.3).

3.2.1 Uniform-scale Vision Transformers

The original Vision Transformer [11] model belongs to this family, where the multi-head self-attention is applied to a consistent scale in the input image where the spatial scale is maintained through the network hierarchy. We name such models as the uniform-scale ViTs, as described below.

Vision Transformer (ViT) [11] (Fig. 6) is the first work to showcase how Transformers can ‘altogether’ replace standard convolutions in deep neural networks on large-scale image datasets. They applied the original Transformer model [1] (with minimal changes) on a sequence of image ‘patches’ flatten as vectors. The model was pre-trained on a large propriety dataset (JFT dataset [47] with 300 million images) and then fine-tuned to downstream recognition benchmarks *e.g.*, ImageNet classification. This is an important step since pre-training ViT on a medium-range dataset would not give competitive results, because the CNNs encode prior knowledge about the images (inductive biases *e.g.*, translation equivariance) that reduces the need of data as compared to Transformers which must discover such information from very large-scale data. Notably, compared to the iGPT [19] model that also applied Transformers to full-sized images but performs training as a generative task, ViT pre-trains the model with a supervised classification task (although a self-supervision variant is also explored which results in a less performance).

The DeiT [12] is the first work to demonstrate that Transformers can be learned on mid-sized datasets (*i.e.*, 1.2 million ImageNet examples compared to 300 million images of JFT [11] used in ViT [11]) in relatively shorter training episodes. Besides using augmentation and regularization procedures common in CNNs, the main contribution of DeiT [12] is a novel native distillation approach for Transformers which uses a CNN as a teacher model (RegNetY-16GF [86]) to train the Transformer model. The outputs from the CNN aid the Transformer in efficiently figuring

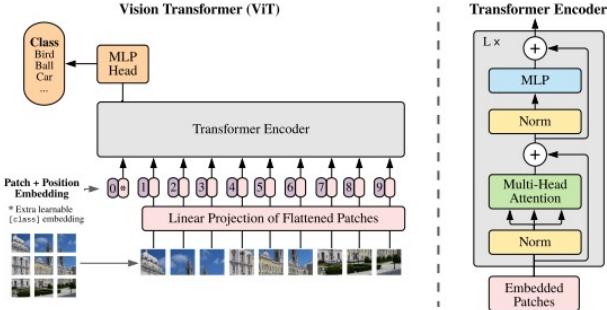


Fig. 6: An overview of Vision Transformer (on the left) and the details of Transformer encoder (on the right). The architecture resembles Transformers used in the NLP domain and the image patches are simply fed to the model after flattening. After training, the feature obtained from the first token position is used for classification. Image obtained from [11].

out useful representations for input images. A distillation token is appended with the input patch embeddings and the class token. The self-attention layers operate on these tokens to learn their inter-dependencies and outputs the learned class, patch, and distillation tokens. The network is trained with a cross-entropy loss defined on the output class token and a distillation loss to match the distillation token with the teacher output. Both *soft* and *hard* label choices were explored for distillation, where the hard distillation was found to perform better. Interestingly, the learned class and distillation tokens do not exhibit a high correlation indicating their complementary nature. The learned representations compare favorably well against top-performing CNN architectures such as EfficientNet [87] and also generalize well for a number of downstream recognition tasks.

Token to Token (T2T) ViT [35] recursively combines neighboring tokens into a single token to reduce tokens length and aggregate spatial context. Transformer in Transformer [88] computes attention at two levels: patch-level (as done in standard ViTs [11]) and local sub-patch-level (e.g. by subdividing a 16×16 patch into four 4×4 blocks, and computing attention amongst these blocks). In token labelling ViT [89], all patch tokens contribute towards loss calculation, different from regular ViTs that only use classification token in the loss. This process includes auxiliary supervision where each image-patch (token) is labeled using a pre-trained CNN model. Similar to CutMix augmentation [90], tokens from different images are mixed as an augmentation strategy, and the model is trained using the standard classification loss and auxiliary token-label loss. Their model demonstrates excellent performance specially for smaller sized models.

The quadratic complexity of self-attention hinders its applicability to longer sequences (high-resolution images). Cross-Covariance Image Transformers (XCiT) [91] incorporate attention across feature-channels instead of tokens, i.e., their cross-covariance attention is given by $\text{Vsoftmax}\left(\frac{\mathbf{K}^T \mathbf{Q}^T}{\sqrt{\tau}}\right)$. The proposed cross-covariance attention has linear complexity (since it depends upon feature dimension instead of the number of tokens). XCiT can therefore handle large resolution images and demonstrate excellent performance across different vision tasks i.e., self-

supervised and fully supervised image classification and dense prediction (detection, segmentation). DeepViT [92] observes that the similarity between attention maps of deeper layer is high and hinders scaling models depth. They propose to re-attend the attention maps in a multi-head block instead of simple aggregation of these attention maps, and show consistent gains over standard multi-head self attention based ViTs.

3.2.2 Multi-scale Vision Transformers

In standard ViTs, the number of the tokens and token feature dimension are kept fixed throughout different blocks of the network. This is limiting, since the model is unable to capture fine spatial details at different scales. Initial Transformer based dense prediction methods (e.g., DETR [13]) therefore have a convolutional backend. Multi-stage hierarchical design for ViTs, where number of tokens is gradually reduced while the token feature dimension is progressively increased, has been shown to produce effective features for dense prediction tasks [36], [93]–[96]. These models generally also perform well for recognition tasks. These architectures mostly sparsify tokens by merging neighboring tokens and projecting them to a higher dimensional feature space. Examples of multi-stage ViTs include Pyramid ViT [93], [97], Twins [37], CoaT [98], Swin Transformer [36], Convolutional vision Transformer (CvT) [96], Shuffle Transformer [95], CrossFormer [99], RegionViT [100] and Focal Transformer models [94]. Some of them are hybrid designs (with both convolution and self-attention operations, see Sec. 3.2.3), while others only employ pure self-attention based design (discussed next).

Pyramid ViT (PVT) [93] is the first hierarchical design for ViT, and proposes a progressive shrinking pyramid and spatial-reduction attention. PVTv2 [97] and SegFormer [101] improve original PVT [93] by introducing overlapping patch embedding, depth-wise convolution, and efficient attention. Swin Transformer [36] has a multi-stage hierarchical architecture which computes attention within a local window, by partitioning the window into multiple sub-patches. To capture interactions between different windows (image locations), window partitioning is gradually shifted, along the hierarchy of the network, to capture overlapping regions. Focal Transformer models [94] is another hierarchical design, where focal self-attention is introduced to simultaneously capture global and local relationships. Similarly, CrossFormer [99] has a hierarchical pyramid structure, and introduces cross-scale embedding module, along-with long short distance attention and dynamic position bias to faithfully capture both local and global visual cues. RegionViT [100] proposes a regional-to-local attention to encode hierarchical features. Multi-Scale Vision Longformer [102] also considers a local context in self-attention, but employs the efficient Longformer [103] design for self-attention. CrossViT [104] encodes multi-scale features with two branches (each with multiple transformer blocks), by separately processeses smaller and larger image patches. The information from these two multi-scale brancnes is then fused together using a cross-attention module.

3.2.3 Hybrid ViTs with Convolutions

Convolutions do an excellent job at capturing low-level local features in images, and have been explored in multiple hybrid ViT designs, specially at the beginning to “patchify and tokenize” an input image. For example, Convolutional vision Transformer (CvT) [96] incorporate convolution based projection to capture the spatial structure and low-level details, for tokenization of image patches. CvT has a hierarchical design, where number of tokens is progressively reduced while the token-width is increased, thus imitating the impact of spatial downsampling as in CNNs. Convolution enhanced image Transformers [105] employ convolutions based image-to-token module to extract low-level features. Compact Convolutional Transformer (CCT) [106] introduces a new sequence pooling scheme, and incorporates convolutional blocks (conv-pool-reshape) for tokenization. CCT can be trained from scratch on smaller datasets, *e.g.*, CIFAR10 with $\sim 95\%$ accuracy, which is a remarkable property not possible with the traditional ViTs.

LocalViT [107] introduces depthwise convolutions to enhance local features modeling capability of ViTs. LeViT [108] (name inspired from LeNet [109]) applies a four-layered CNN block (with 3×3 convolutions) at the beginning with progressively increasing channels (3,32,64,128,256). For a $3 \times 224 \times 224$ input image, the resulting $256 \times 14 \times 14$ output from the CNN block becomes input to a hierarchical ViT. By virtue of its design, LeViT is $5\times$ faster than EfficientNet [87] on CPU, at inference. ResT [110] is another hierarchical architecture which applies a CNN block at the beginning for patch-embedding. It incorporates depth-wise convolutions and adaptive position encoding to tackle varying image sizes. A recent approach NesT [111] proposes a simple technique to introduce hierarchy in ViTs. NesT divides an image into non-overlapping blocks (each block is further split into patches). It first separately applies local self-attention on patches within each block, and then enables global interaction between blocks by aggregating them into an image space and applying convolution operation, followed by downsampling. The number of blocks is gradually reduced along the hierarchy of the model, while number of local-patches is kept fixed. This simple scheme performs favorably compared with more sophisticated designs [36], [97], and enables training NesT on smaller datasets (*e.g.*, CIFAR-10) from scratch.

Depthwise Convolution and self-Attention Networks (CoAtNets) [112] introduce a relative attention module (which combines depthwise convolutions and self-attention), and vertically stack convolution and attention layers. CoAtNets demonstrate an impressive 86% ImageNet top-1 accuracy without extra data (*i.e.* trained only on ImageNet-1k). Shuffle Transformer [95] performs self-attention within a window and has depth-wise convolutions between the window-based multi-head self-attention and MLP. It introduces a shuffle operation to build stronger cross-patch connections. Co-scale conv-attentional image Transformers (CoAT) [98], is a hybrid hierarchical pyramid design, with serial and parallel blocks, where the serial block is similar to standard transformer block except for the attention layer replaced with depthwise convolution. The parallel blocks is applied on the output of serial blocks

and encodes relationships between tokens at multiple scales using cross-attention. Twins [37] builds upon PVT [93] (an attention only pyramid design), by replacing the absolute position embedding in PVT with relative conditional position embedding [113], and incorporating the separable depth-wise convolutions instead of the standard spatial attention, to capture local and global context of the image. In this sense, the hybrid designs tend to combine the strengths of both convolution and transformer models. TransCNN [114] propose a hierarchical multi-head self attention block, which first learns interactions within small grids (tokens) using self-attention, and then gradually merges the smaller grids into larger grids. The proposed block can then be plugged into existing CNN architectures.

3.2.4 Self-Supervised Vision Transformers

Contrastive learning based self-supervised approaches, which have gained significant success for CNN based vision tasks, have also been investigated for ViTs. Chen *et al.* [115] evaluate different self-supervised frameworks and propose practical strategies including MoCo v3 (extended from v1/v2 [116], [117]) for stabilized training of self-supervised ViTs. Xie *et al.* [118] combine MoCo v2 [117] and BYOL [119] to train DeiT [12] and SwinTransformer [36]. They demonstrate generalization of self-supervised SwinTransformer for dense prediction tasks of detection and segmentation. Self distillation with no labels (DINO) [120] demonstrate that self-supervised ViTs can automatically segment the background pixels of an image, even though they were never trained using pixel-level supervision, a phenomena otherwise not observed in CNNs or fully supervised ViTs. Efficient self-supervised vision transformer (EsViT) [121] propose a multi-stage design, where neighboring tokens are gradually merged along the hierarchy of the network, and use DINO for self-supervision. Apart from standard image-level self-supervision as in DINO, they incorporate additional patch-level self-supervision in which correspondence is promoted between similar patches within augmented versions of an image. EsViT demonstrates excellent performance under self-supervision settings, and its off-the-shelf features transfer better than supervised SwinTransformer on 17 out of 18 evaluated datasets.

3.3 Transformers for Object Detection

Transformers based modules have been used for object detection in the following manner: (a) Transformer backbones for feature extraction, with a R-CNN based head for detection (see Sec. 3.2.2), (b) CNN backbone for visual features and a Transformer based decoder for object detection [13], [14], [122], [123] (see Sec. 3.3.1, and (c) a purely transformer based design for end-to-end object detection [124] (see Sec. 3.3.2).

3.3.1 Detection Transformers with CNN Backbone

Detection Transformer (DETR) [13] treats object detection as a set prediction task *i.e.*, given a set of image features, the objective is to predict the set of object bounding boxes. The Transformer model enables the prediction of a set of objects (in a single shot) and also allows modeling their relationships. DETR adapts a set loss function which allows

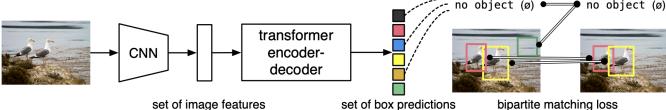


Fig. 7: Detection Transformer (DETR) [13] treats the object detection task as a set prediction problem and uses the Transformer network to encode relationships between set elements. A bipartite set loss is used to uniquely match the box predictions with the ground-truth boxes (shown on the right two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

bipartite matching between predictions and ground-truth boxes. The main advantage of DETR is that it removes the dependence on hand-crafted modules and operations, such as the RPN (region proposal network) and NMS (non-maximal suppression) commonly used in object detection [125]–[129]. In this manner, the dependence on prior knowledge and careful engineering design is relaxed for complex structured tasks like object detection.

Given spatial feature maps from the CNN backbone, the encoder first flattens the spatial dimensions (see Fig. 7). This gives a sequence of features $d \times n$, where d is the feature dimension and $n = h \times w$ with h, w being the height and width of the spatial feature maps. These features are then encoded and decoded using multi-head self-attention modules as in [1]. The main difference in the decoding stage is that all boxes are predicted in parallel while [1] uses an RNN to predict sequence elements one by one. Since the encoder and decoder are permutation invariant, learned positional encodings are used as the object queries by the decoder to generate different boxes. Note that the spatial structure in a CNN detector (e.g., Faster R-CNN) automatically encodes the positional information. DETR obtains performance comparable to the popular Faster R-CNN model [125] which is an impressive feat given its simple design. The DETR has also been extended to interesting applications in other domains, e.g., Cell-DETR [130] extends it for instance segmentation of biological cells. A dedicated attention branch is added to obtain instance-wise segmentations in addition box predictions that are enhanced with a CNN decoder to generate accurate instance masks.

The DETR [13] model successfully combines convolutional networks with Transformers [1] to remove hand-crafted design requirements and achieves an end-to-end trainable object detection pipeline. However, it struggles to detect small objects and suffers from slow convergence and a relatively high computational cost [14]. DETR maps images to features space before using the Transformer for the relation modeling. Thus, the computational cost of self-attention grows quadratically with the spatial size of the feature map i.e., $\mathcal{O}(H^2W^2C)$, where H and W represent the height and width of the feature map. This inherently puts a limitation on the use of multi-scale hierarchical features [131] in DETR training framework which is ultimately important to detect small objects. Furthermore, at the beginning of training, the attention module simply projects uniform attention to all the locations of the feature map and

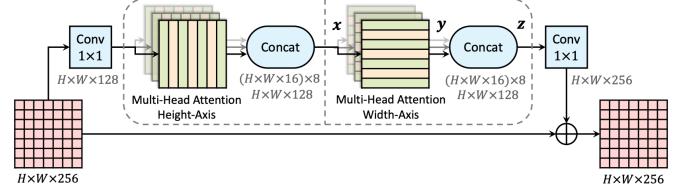


Fig. 8: Axial attention module [133] that sequentially applies multi-head axial attention operations along height and width axes. Image from [133].

requires a large number of training epochs to tune attention weights to converge to meaningfully sparse locations. This approach contributes to a slow convergence rate of DETR. To mitigate the above-mentioned issues, [14] proposed a deformable attention module to process the feature maps. Inspired from deformable convolutions [42], deformable attention module [14] only attends to sparse set of elements from the whole feature map regardless of its spatial size. This further allows cross-scale aggregation of feature maps with the help of multi-scale attention modules without increasing the computational cost significantly. Deformable DETR not only performs better but its training time also remains $10\times$ lower than the original DETR model [14]. Anchor DETR [122] replaces the learnable query tokens in [13] with anchor-point based queries, such that each query focuses on predicting the object near the anchor point. The anchor points can be fixed on 2D grid, or learned from uniformly distributed points. Anchor DETR [122] requires $10\times$ fewer training epochs with comparable performance. Pix2Seq [123] is a generic Transformer-based framework, without any specialized task-specific modules, and learns to directly produce a sequence of tokens with object descriptions (bounding-boxes and class-labels). A quantization and serialization scheme first converts bounding boxes and class-labels into a sequence of discrete tokens. A generic Transformer based encoder-decoder network is then used to generate these tokens in an auto-regressive manner conditioned on previous predictions and image features.

3.3.2 Detection with Pure Transformers

You Only Look at One Sequence (YOLOS) [124] is a simple, attention-only architecture directly built upon the ViT [1], [132]. It replaces the class-token in ViT with multiple learnable object query tokens, and the bipartite matching loss is used for object detection similar to [13]. YOLOS demonstrates the flexibility of ViTs to object detection, in a pure sequence-to-sequence learning manner, with minimal image related 2D inductive biases. In similar spirit, PVT [93] is combined with DETR [13] to perform object detection with an end-to-end transformer pipeline. We note that it is feasible to combine other recent ViTs with transformer based detection heads as well to create pure ViT based designs [124], and we hope to see more such efforts in future.

3.4 Transformers for Segmentation

Self-attention can be leveraged for dense prediction tasks like image segmentation that requires modeling rich interactions between pixels. Below, we discuss axial self-attention

[133], a cross-modal approach [15] that can segment regions corresponding to a given language expression, and ViTs based segmentation architectures [101], [134], [135].

Panoptic segmentation [136] aims to jointly solve the otherwise distinct tasks of semantic segmentation and instance segmentation by assigning each pixel a semantic label and an instance id. Global context can provide useful cues to deal with such a complex visual understanding task. Self-attention is effective at modeling long-range contextual information, albeit applying it to large inputs for a dense prediction task like panoptic segmentation is prohibitively expensive. A naive solution is to apply self-attention either to downsampled inputs or to limited regions around each pixel [81]. Even after introducing these constraints, the self-attention still has quadratic complexity and sacrifices the global context. To tackle these issues, Wang *et al.* [133] propose the position-sensitive axial-attention where the 2D self-attention mechanism is reformulated as two 1D axial-attention layers, applied to height-axis and width-axis sequentially (see Fig. 8). The axial-attention is compute efficient and enables models to capture the full-image context. It achieves competitive performance for the panoptic segmentation task on COCO [75], Mapillary Vistas [137], and Cityscapes [73] benchmarks and for the image classification on ImageNet dataset [138].

Cross-modal Self-attention (CMSA) [15] encodes long-range multi-modal dependencies between linguistic and visual features for *referring image segmentation task*, that aims to segment entities in an image referred by a language description. For this purpose, a set of cross-modal features is obtained by concatenating image features with each word embedding and the spatial coordinate features. The self-attention operates on these features and generates attention over the image corresponding to each word in the sentence. The segmentation network then performs self-attention at multiple spatial levels and uses a gated multi-level fusion module to refine segmentation masks via information exchange across multi-resolution features. A binary CE loss is used to train the overall model that achieves good improvements on UNC [139], G-Ref [140] and ReferIt [141] datasets.

While the segmentation approaches discussed above insert self-attention in their CNN based architectures, some recent works have proposed transformer based encoder-decoder architectures. Segmentation Transformer (SETR) [134] has a ViT encoder, and two decoder designs based upon progressive upsampling, and multi-level feature aggregation. SegFormer [101] has a hierarchical pyramid ViT [93] (without position encoding) as an encoder, and a simple MLP based decoder with upsampling operation to get the segmentation mask. Segmenter [135] uses ViT encoder to extract image features, and the decoder is a mask Transformer module which predicts segmentation masks, using learnable mask tokens and image-patch tokens as inputs. The authors also propose a baseline linear decoder which projects the patch-embeddings to classification space, thus producing coarse patch-level labels.

3.5 Transformers for Image and Scene Generation

Here, we discuss Transformer-based architectures [23], [142]–[146] for image synthesis, which is interesting from

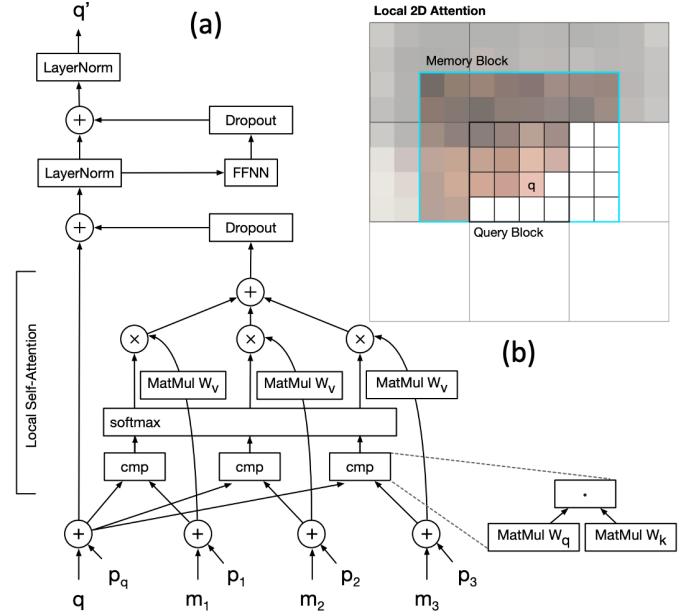


Fig. 9: (a) Self-attention block in Image Transformer [142]. Given one channel for a pixel q , the block attends to the memory of previous synthesized pixels (m_i), followed by a feed-forward sub-network. Positional encodings p_i are added in the first layer. (b) The operation performed in Local Self-Attention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

the perspective of generative modeling and learning unsupervised representations for down-stream tasks.

Parmar *et al.* [142] develop an image generation model that can sequentially predict each pixel of an output image given its previously generated pixels (Fig. 9). Their approach models the joint distribution of the image pixels by factorizing it as a product of pixel-wise conditional distributions. Previously developed auto-regressive models for this task, such as the PixelCNN [147], suffer from a limited receptive field which hinders in modeling long term relationships in an image *e.g.*, part relationships or occlusions. Using self-attention, [142] enhances the receptive field without incurring a high computational cost (*e.g.*, effective receptive field up to 256 pixels can be achieved as compared to 25 pixels of PixelCNN [147]). The generative pipeline was also tested on conditional generation tasks *e.g.*, image super-resolution, image completion, and denoising.

Inspired by the success of GPT model [5] in the language domain, image GPT (iGPT) [143] demonstrated that such models can be directly used for image generation tasks, and to learn strong features for downstream vision tasks (*e.g.*, image classification). Specifically, iGPT trains GPT v2 model [5] on flattened image sequences (1D pixel arrays) and shows that it can generate plausible image outputs without any external supervision. The generated samples depict the model's ability to understand spatial relationships between pixels and high-level attributes such as object classes, texture, and scale. Notably, the design

does not use any image-specific knowledge in the design (e.g., the 2D position embeddings used in Image Transformer [142]). The features learned with iGPT’s unsupervised training mechanism compete impressively against other unsupervised approaches, achieving state-of-the-art performance on CIFAR-10/100 [148] and STL [149] datasets while performing comparably to SimCLR (a contrastive learning approach) [150] on ImageNet dataset. This is an astounding result, since the iGPT architecture is exactly the same as used for language modeling tasks, and therefore it does not incorporate any prior domain-specific knowledge. Notably, the competing unsupervised CNN based solutions widely adopt such priors in the form of architectural design, attention mechanisms, loss functions, and regularization [117], [151]–[154]. However, on the downside, iGPT has a high compute cost e.g., iGPT-L version has roughly $36 \times$ high training cost compared to MoCo [117] which is a state of the art self-supervised feature learning approach. For this reason, the training was generally limited to low-resolution of $\leq 64 \times 64$, while convolutional architectures can effectively learn from high-resolution inputs.

Transformers typically incur a high compute cost when applied on high-dimensional sequences. To overcome this limitation, Esser *et al.* [144] proposed to include inductive biases (commonly used in the CNNs) alongside Transformers to improve their efficiency. Specifically, local connectivity and spatial invariance biases inbuilt in the CNN structure are leveraged by learning a rich dictionary of visual patterns (using a Generative Adversarial approach). A Transformer is then used to learn the long-range interactions between the dictionary items to generate the outputs. In turn, they develop a conditional image generation model capable of producing very high-resolution images (up to megapixel range) using Transformers. This is the first work that demonstrates the application of Transformers to generate such high-resolution images.

Generative Adversarial Networks (GANs) [54] with CNNs as default backbone have been very successful for visually appealing image synthesis [155]–[157]. TransGAN [145] builds a strong GAN model, free of any convolution operation, with both generator and discriminator based upon the Transformer model [1]. The architecture of both generator and discriminator is based upon the encoder in original Transformer model [1]. For memory efficiency, the generator contains multiple stages, with up-sampling modules in-between, which gradually increase the resolution of feature maps (input sequence length) while reducing the embedding dimension. The discriminator of TransGAN takes flattened image-patches as tokens similar to [132]. Authors introduce different training techniques including data augmentation, training with an auxiliary task and injecting locality to self-attention to scale-up their model for high quality image synthesis [144]. The TransGAN model achieves state-of-the-art results in terms of Inception Score and Fréchet Inception Distance (FID) on STL-10 and performs favorably compared with their CNN-based GAN counterparts on other datasets.

Unlike previous image generation methods [142]–[144], which directly predict image outputs, [23] learns to generate parameters of 3D objects to be placed in a given scene. Specifically, SceneFormer [23] studies the 3D room layout

conditioned scene generation task. Given the empty room shape, [23] can propose new object configurations in the room while maintaining realism. Remarkably, the model does not use any appearance information and only learns to generate new scenes by modeling the inter-object relationships using self-attention in Transformers. Similar to how a Transformer operates on a sentence, it is applied to a sequence of objects to predict the next suitable object in a scene. Specifically, the size, pose, location, and category of the next object is predicted by the Transformer model. A start token indicates the initiation of inference and the number of output token indicate the objects generated by the model in a sequence. The authors also explore generating new scenes given a textual description of the room layout. The independence from the appearance makes the approach efficient, enabling interactive scene generation.

The task of generating realistic images from text is interesting and practically valuable (e.g., for artistic content creation), but at the same time highly challenging. Prior text-to-image synthesis approaches [158]–[161] are mostly based on GANs [54]. Although these methods produce encouraging results, they are far from being photo-realistic. Ramesh *et al.* [20] recently proposed DALL-E which is a Transformer model capable of generating high-fidelity images from a given text description. DALL-E model has 12 billion parameters and it is trained on a large set of text-image pairs taken from the internet. Before training, images are first resized to 256×256 resolution, and subsequently compressed to a 32×32 grid of latent codes using a pre-trained discrete variational autoencoder [162], [163]. DALL-E takes as input a single stream of 1280 tokens (256 for the text and 1024 for the image), and is trained to generate all other tokens autoregressively (one after another). It provides flexibility to generate images either from scratch (Fig. 10a) or by extending existing images (Fig. 10b), while staying faithful to the text caption.

The authors demonstrate the effectiveness of DALL-E by creating images from text describing a wide variety of real and fictional concepts. While generating images purely from textural captions, DALL-E shows impressive performance at controlling multiple objects and their attributes (Fig. 10c), rendering certain viewpoint (Fig. 10d), capturing object’s internal structure (Fig. 10e), and combining unrelated objects (Fig. 10f). Furthermore, DALL-E can perform image-to-image translation (Fig. 10g) guided by the input text.

3.6 Transformers for Low-level Vision

After witnessing the success of Transformer models in high-level vision problems, numerous Transformer-based methods have been proposed for low-level vision tasks, including image super-resolution [16], [19], [164], denoising [19], [165], deraining [19], [165], and colorization [24]. Image restoration requires pixel-to-pixel correspondence from the input to the output images. One major goal of restoration algorithms is to preserve desired fine image details (such as edges and texture) in the restored images. CNNs achieve this by employing a single-scale architecture design that does not involve any downsampling operation. Since the computational complexity of self-attention in Transformer models increases quadratically with number of image patches, it is

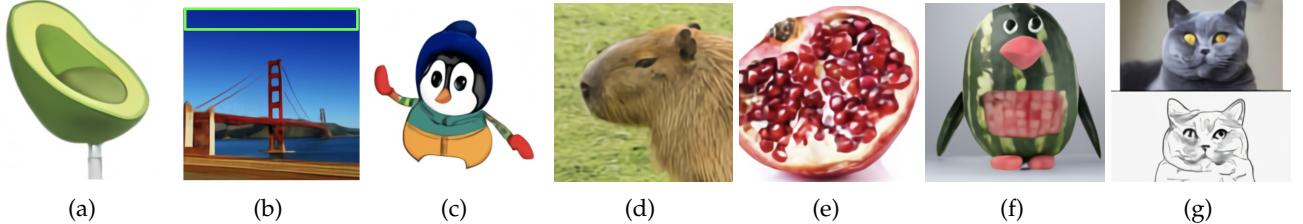


Fig. 10: Images generated by DALL-E [20] from the following text prompts. (a) *An armchair in the shape of an avocado.* (b) *A photo of San Francisco's golden gate bridge.* Given a part of the image (in green box), DALL-E performs the image completion. (c) *An emoji of a baby penguin wearing a blue hat,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

infeasible to develop Transformer model that can operate on single-scale feature processing pipeline. Consequently, these Transformer-based image restoration models make use of various strategies to reduce the computational burden, such as computing attention on local image windows [164], performing spatial reduction attention [166], and employing encoder-decoder design [19], [165]. Here, we briefly discuss a few image restoration Transformer models.

3.6.1 Transformers for Image Processing Tasks

Top performing algorithms for high-level computer vision tasks such as object detection and semantic segmentation often employ backbone models that are pre-trained on large-scale datasets *e.g.*, ImageNet. In contrast, algorithms for low-level vision tasks such as image denoising, super-resolution, and deraining are directly trained on task-specific data, thereby suffer from these limitations: (i) small number of images available in task-specific datasets (*e.g.*, the commonly used DIV2K dataset for image super-resolution contains only 2000 images), (ii) the model trained for one image processing task does not adapt well to other related tasks.

Chen *et al.* [19] propose a pre-trained model based on Transformer architecture, named as Image Processing Transformer (IPT). It is capable of performing various image restoration tasks such as super-resolution, denoising, and deraining. The overall architecture of IPT consists of multi-heads and multi-tails to deal with different tasks separately, and a shared encoder-decoder Transformer body. Since exploiting Transformers at full potential requires training on large-scale data, [19] takes the clean (ground-truth) images from the ImageNet benchmark and synthesize their degraded versions for different tasks. For example, bicubic interpolation is used for generating low-resolution images, additive white Gaussian noise is added to prepare noisy data, and hand-crafted rain streaks are applied to obtain rainy images. In total, 10 million images are used to pre-train the IPT model. During training, each task-specific head takes as input a degraded image and generates visual features. These feature maps are divided into small crops and subsequently flattened before feeding them to the Transformer encoder (whose architecture is the same as [1]). The outputs of the encoder along with the task-specific embeddings are given as input to the Transformer decoder. The features from the decoder output are reshaped and passed to the multi-tail that yields restored images. The IPT model is optimized with L_1 loss. Experimental results show that the pre-trained IPT model, when fine-tuned for a specific low-level vision

task, can provide significant performance gains over the state-of-the-art methods [167]–[169].

3.6.2 Transformers for Super-Resolution

Recent years have seen major performance breakthroughs for super-resolution (SR) due to convolutional neural networks (CNNs). Principally, the quality of super-resolved images generated by CNNs is dependent on the choice of optimization objective. While the SR methods [167], [170]–[173] that are based on pixel-wise loss functions (*e.g.*, L_1 , MSE, etc.) yield impressive results in terms of image fidelity metrics such as PSNR and SSIM, they struggle to recover fine texture details and often produce images that are overly-smooth and perceptually less pleasant. Further, *perceptual* SR approaches [52], [174]–[177], in addition to per-pixel loss, employ adversarial loss [54] and perceptual loss [178] based on deep features extracted from pre-trained CNNs. While these methods generate images that are sharp, visually pleasant, and perceptually plausible, they show a substantial decrease in reconstruction accuracy measured in PSNR/SSIM. Moreover, the perceptual SR algorithms have a tendency to hallucinate fake textures and cause artifacts. The above mentioned SR approaches follow two distinct (but conflicting) research directions: one maximizing the reconstruction accuracy and the other maximizing the perceptual quality, but never both.

To alleviate the trade-off between perceptual reproduction and accurate reproduction, Yang *et al.* [16] propose a Transformer network (TTSR) for super-resolution. During training, TTSR uses paired LR-HR images, as well as reference (Ref) images with similar content as of LR images. TTSR learns to search relevant regions in the Ref image and transfers rich textures to help super-resolving the input LR image. The texture Transformer module of TTSR method (see Fig. 11) consists of four core components: (1) *Learnable texture extractor*: takes as input $LR\uparrow$, $Ref\downarrow\uparrow$, and Ref images, and generates texture features query (Q), key (K), and value (V), respectively. Here, \uparrow denotes bicubic upsampling operation, and $\downarrow\uparrow$ represents bicubic down-sampling followed by an upsampling operation. (2) *Relevance embedding*: first unfolds Q and K into patches and then computes the similarity of each patch in Q with each patch in K in order to generate hard and soft attention maps. (3) *Hard-attention*: transfers HR texture features from V to (LR features) Q using the hard attention map. (4) *Soft-attention*: further enhances relevant features while suppressing less relevant ones.

While TTSR [16] method deals with reference-based image super-resolution, most of the research is conducted

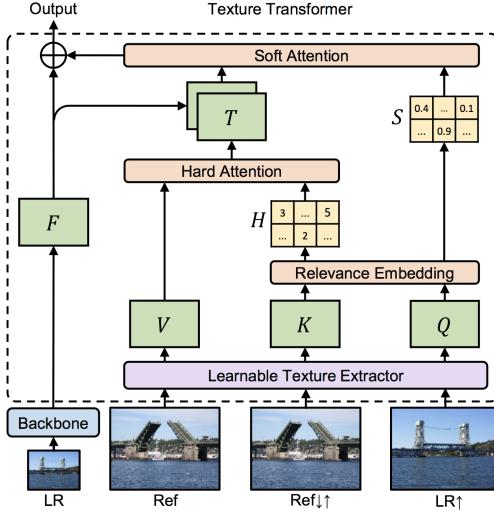


Fig. 11: Diagram of the texture Transformer module. Q (query), K (key) and V (value) represent texture features extracted from a (bicubic upsampled) low-resolution image, a sequentially down/upsampled reference image, and an original reference image, respectively. The relevance embedding aims to estimate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

on single image super-resolution problem in which only LR-HR paired images are available. Since the computational complexity of the original self-attention operation is prohibitively high for high-resolution images, recently a few efficient transformer models have been proposed that employ window-based attention (SwinIR [164]) and spatial resolution reduction operation in attention module (ESRT [166]) to perform super-resolution.

3.6.3 Colorization Transformer

Given a grayscale image, colorization seeks to produce the corresponding colorized sample. It is a one-to-many task as for a given grayscale input, there exist many possibilities in the colorized output space. The challenging nature of this task requires probabilistic models capable of producing multiple colorized output samples. Colorization Transformer [24] is a probabilistic model based on conditional attention mechanism [179]. It divides the image colorization task into three sub-problems and proposes to solve each task sequentially by a different Transformer network. The authors first train a Transformer network to map a low-resolution grayscale image to a 3-bit low-resolution colored image. Low-resolution images in turn allow training of larger models. The 3-bit low-resolution colored image is then upsampled to an 8-bit RGB sample by another Transformer network in the second stage of training. Finally, a third stage Transformer is trained to increase the spatial resolution of the 8-bit RGB sample produced by the second-stage Transformer. Self-attention used in the colorization Transformer is based on row/column attention layers introduced in [179]. These layers capture the interaction between each pixel of an input image while being computationally less costly. The row-wise attention layer applies self-

attention to all pixels in a given row, while the column-wise attention layer considers pixels only in a given column of an image. This work [24] is the first successful application of Transformers trained to colorize grey-scale images at high (256×256) resolution.

3.7 Transformers for Multi-Modal Tasks

Transformer models have also been extensively used for vision-language tasks such as visual question answering (VQA) [183], visual commonsense reasoning (VSR) [184], cross-modal retrieval [185] and image captioning [29]. Several works in this direction target effective vision-language pre-training (VLP) on large-scale multi-modal datasets to learn generic representations that effectively encode cross-modality relationships (e.g., grounding semantic attributes of a person in a given image). These representations can then be transferred to downstream tasks, often obtaining state of the art results. Notably, several of these models still use CNNs as vision backbone to extract visual features while Transformers are used mainly used to encode text followed by the fusion of language and visual features. Such models generally apply the vanilla multi-layer Transformer [1] with multi-modal inputs and do not introduce fundamental changes to the core attention block. However, their main distinction is in the configuration of Transformers and the loss functions, based on which we categorize them into: (a) Multi-stream Transformers (see Sec. 3.7.1) and (b) Single-stream Transformers (see Sec. 3.7.2). The *single-stream* designs feed the *multi-modal* inputs to a single Transformer while the *multi-stream* designs first use independent Transformers for each modality and later learn cross-modal representations using another Transformer (see Fig. 12). Besides these vision language pretraining methods, we also explain visual grounding approaches towards the end of this section (see Sec. 3.7.3).

3.7.1 Multi-stream Transformers

Vision and Language BERT (ViLBERT) [63] was the first extension of the BERT model to the multi-modal domain. The goal was to learn representations that can jointly model images and natural language. For this purpose, ViLBERT developed a two-stream architecture where each stream is dedicated to model the vision or language inputs (Fig. 12-h). The architecture of both parallel streams is a series of Transformer blocks similar to the BERT model. Subsequently, co-attentional Transformer layers are applied to learn cross-modal relationships. The co-attentional framework is very simple. Query, key, and value matrices are computed for each modality in the standard way [1] and then key-value pairs for one modality are passed on to the other modality's attention head.

ViLBERT applies VLP on a set of proxy tasks defined on the Conceptual Concepts dataset (with 3.3M images with weak captions) and later fine-tune the model on downstream tasks such as VQA. The pre-training phase operates in a self-supervised manner, i.e., pretext tasks are created without manual labeling on the large-scale unlabelled dataset. These pretext tasks include predicting whether the text and image inputs are related and predicting the semantics of masked image regions and textual inputs (e.g., similar

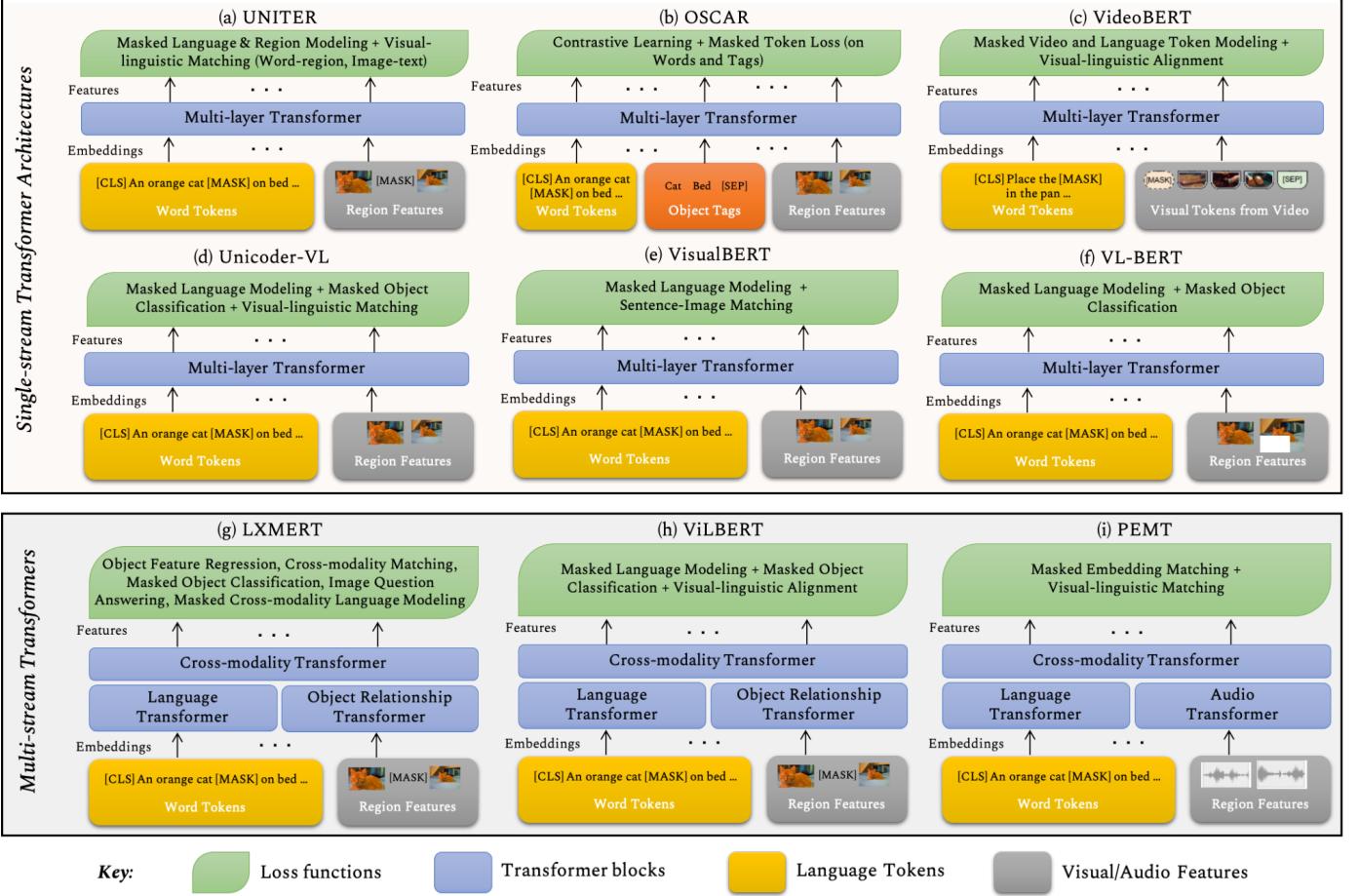


Fig. 12: An overview of Transformer models used for multi-modal tasks in computer vision. The Transformer designs in this category can be grouped into single-stream (UNITER [43], OSCAR [44], VideoBERT [17], Unicoder-VL [180], VisualBERT [63] and VL-BERT [22]) and dual-stream architectures (LXMERT [21], ViLBERT [181] and PEMT [182]). A key distinction between models is the choice of loss functions. While most of the multi-modal methods are focused on images as visual data, VideoBERT [17] and PEMT [182] are designed to work on video streams and leverage unique modalities e.g., audio signals in videos [182].

to reconstructing masked words in text in the BERT model [3]). This way, the model learns the inherent structure in the data during pre-training and also models cross-domain associations. With evaluations on several tasks, [17] demonstrated that a two-stream model can perform better than a single-stream model that uses shared parameters to model both language and vision domains [17].

Similar to ViLBERT [181], Learning Cross-Modality Encoder Representations from Transformers (LXMERT) [21] also uses a two-stream architecture based on BERT framework. The main difference lies in the object-relationship encoder that is used to model the visual features instead of simple image-level features used in ViLBERT. The information in two streams is then fused across modalities using cross-attention blocks similar to [181].

Compared to two pre-texts tasks used for VLP in [181], LXMERT uses five pre-training tasks including masked object and language prediction, cross-modality matching, and visual question answering (Fig. 12-g). The pre-trained model is fine-tuned on the VQA task, however, a high similarity between pre-training and fine-tuned tasks raises questions on the generalizability of the learned representations to new tasks. To this end, the authors conducted generalization

experiments on Visual Reasoning for Real (NLVR) task [186] demonstrating impressive improvements on novel tasks.

Lee *et al.* [182] note that the multi-modal representation learning approaches like VideoBERT [17] and ViLBERT [181] generally keep the language processing part fixed to a pre-trained model (e.g., BERT [3]) to reduce training complexity. For the first time in the literature, they propose to learn an end-to-end multi-modal bidirectional Transformer model called PEMT on audio-visual data from unlabeled videos. First, short-term (e.g., 1-3 seconds) video dynamics are encoded using CNNs, followed by a modality-specific Transformer (audio/visual) to model long-term dependencies (e.g., 30 seconds). A multi-modal Transformer is then applied to the modality-specific Transformer outputs to exchange information across visual-linguistic domains. However, learning such a model in a naive form would incur huge memory requirements. To reduce parametric complexity, the parameters are shared across layers within each Transformer which leads upto 80% parameter reduction. The Transformer is trained using a contrastive learning approach based on a content-aware negative sampling (Fig. 12-i). Specifically, the model uses the features obtained from CNNs learned during the training phase to select negative

samples that are visually similar to the positive instances. This work also compares various fusion strategies adopted in earlier works such as early (VideoBERT [17] and VL-BERT [22]), mid-level (ViL-BERT [181] and LXMERT [21]) and late fusion mechanisms and shows that the mid-level fusion is the optimal choice. The proposed model is pre-trained on Kinetics-700 [187] dataset and later fine-tuned on downstream video classification tasks such as short video classification on UCF101 [188], audio classification on ESC50 [189] and long-term action recognition on Charades [190] and Kinetics-Sounds [65] datasets.

Tan and Bansal [191] introduce the concept of ‘vokens’ (images related to language tokens extracted from sentences). The vokens (visualized tokens) provide visual supervision to the language model to learn better features. The motivation is that humans learn languages by correlating visual information with semantic concepts. In a similar spirit to other self-supervised language representation learning methods [3], [181], they learn representations by defining an auxiliary task of voken-prediction task. Since the existing datasets encode limited visually grounded tokens, they propose a vokenization method to map language tokens to visual vokens, as illustrated in Fig. 13. The approach uses language-based retrieval for such a mapping and transfers a model trained on a small labeled dataset (MS-COCO) to a large dataset (Wikipedia). Furthermore, it was ensured that the sentence-wide context is considered to obtain the token-voken mapping. The resulting model trained using generated tokens outperforms the state of the art BERT model on a diverse set of NLP tasks. In this sense, the proposed model does not evaluate vision tasks, however, uses vision as a useful grounding cue to train the language model, hence we include it in the multi-modal representation learning group.

Vision-and-Language Navigation (VLN) aims to predict a navigation plan on a map based on the vision and language inputs. Transformer models were used earlier in [192], [193] for VLN task. These works first pre-train a cross-modal Transformer using self-supervision on vision and language pairs and subsequently fine-tune on the specific VLN tasks. While these works learn attention between image region and language, Chen *et al.* [194] propose to learn cross-modal attention between language inputs and spatial topological maps (to represent an agent’s environment as a graph whose nodes denote places and the edges denote their connectivity). Given the topological map and natural language inputs, a VLN task using the Transformer model bears resemblance to sequence prediction in NLP. Specifically, at each time instance, the cross-modal Transformer predicts a single node of the topological map in the navigation plan. The individual language and map encodings are first processed using uni-modal encoders and later a cross-modal encoder (similar to LXMERT [21]) is applied to aggregate information across modalities. To denote positions in the map, a learned trajectory position encoding is appended with the map features. Based on this Transformer setup, [194] reports a full navigation system that can freely explore the environment and intelligently plan its actions.

CLIP [195] is a contrastive approach to learn image representations from text, with a learning objective which maximizes similarity of correct text-image pairs embeddings in a large batch size. Specifically, given a batch of N image-

text pairs, CLIP learns a multi-modal embedding space, by jointly training an image-encoder and a text-encoder, such that the cosine similarity of the valid N image-text pairs is maximized, while the remaining $N^2 - N$ pairs is minimized. The authors consider ResNet-50 [67] and Vision Transformer (ViT) [132] for encoding images. The modified Transformer model [1] as in [5] is employed for encoding text. CLIP is trained on a large corpus of 400 million image-text pairs and demonstrates excellent zero-shot transfer capabilities. At inference, the names of classes are used as input to the text-encoder, and similarity of the encoded image is computed with all encoded texts (classes) to find the image-text pair with highest match. The CLIP achieves an astounding zero-shot classification accuracy of 75% on ImageNet, without using any supervision from ImageNet training set. The authors further demonstrate zero-shot transfer capabilities of the CLIP model on 30 different computer vision benchmarks. Note that CLIP with ResNet took 18 days to train on 592 V100 GPUs while CLIP with ViT took 12 days on 256 V100 GPUs. This highlights the computational cost of CLIP.

3.7.2 Single-stream Transformers

Different from two-stream networks like ViLBERT [181] and LXMERT [21], VisualBERT [63] uses a single stack of Transformers to model both the domains (images and text). The input sequence of text (*e.g.*, caption) and the visual features corresponding to the object proposals are fed to the Transformer that automatically discovers relations between the two domains. Notably, VisualBERT architecture is somewhat similar to VideoBERT [17] (explained in Sec. 3.8), but instead of only focusing on cooking videos, VisualBERT evaluates on various visual-linguistic tasks (*e.g.*, VCR, NLVR, VQA, and visual grounding). The VisualBERT model first applies task-agnostic pre-training using two objectives (Fig. 12-e). The first objective simply attempts to predict missing text tokens using the image features and remaining textual tokens. The second objective attempts to differentiate between the true and false caption of a given image. After task-agnostic pre-training, the authors propose to perform task-specific pre-training to bridge the domain gap before the final fine-tuning to the downstream task.

Su *et al.* [22] propose a multi-modal pre-training approach to learn features that are generalizable to multi-modal downstream tasks such as Visual Commonsense Reasoning and Visual Question Answering. This endeavor requires adequately aligning the visual and linguistic cues so that an effective composite representation is learned. To the end, [22] builds on the BERT model and inputs both the visual and language features. The language features correspond to the token in the input sentence and the visual features correspond to the region of interest (RoI) from the input image (obtained via a standard Faster R-CNN). Specifically, the model is pre-trained on both the visual-lingual dataset (Conceptual Captions [196]) as well as the language-only datasets (*e.g.*, Wikipedia). The loss function is identical to BERT, where the model is trained to predict the masked out words or visual ROIs (Fig. 12-f). In contrary to other works such as UNITER [43], VL-BERT claims that the visual-linguistic matching tasks are not useful during pre-training, which is in contrast to evidence from later efforts

[180]. Their results on several multi-modal tasks show their benefit over the language-only pre-training (*e.g.*, in BERT).

Universal Encoder for Vision and Language (Unicoder-VL) [180] learns multi-modal representations using large-scale image-caption pairs. The language and image inputs are fed to a single Transformer model (with multiple successive encoders) to learn joint embeddings. To this end, it uses masked word prediction, masked object classification, and visual-linguistic matching as self-supervision tasks during pre-training (Fig. 12-d). Notably, the visual-linguistic matching is carried out only at the global level (*i.e.*, image-sentence alignment). The model is evaluated on image-text retrieval, zero-shot learning, and visual commonsense reasoning where it performs better than the previous models such as ViLBERT [181] and VisualBERT [63]. This shows the significance of rich self-supervised tasks and advocates for a unified Transformer architecture to learn multi-modal features in a common framework.

The Unified Vision-Language Pre-training (VLP) [197] model uses a single Transformer network for both encoding and decoding stages. This stands in contrast to BERT inspired VLP models [17], [22], [63], [198] which use independent encoder and decoder networks. Joint modeling of encoding and decoding stages allows the Unified VLP model to perform well for both image captioning and visual-question answering tasks, when fine-tuned on these individual tasks. The intuition for shared modeling of encoding and decoding stage stems from the need to better share cross-task information during pre-training. The unified model consists of a stack of 12 Transformer blocks, each with a self-attention layer followed by a feed-forward module. The self-supervised objectives used for pre-training include masked vision-language predictions. Here, the authors explore two variants *i.e.*, bidirectional and sequence-to-sequence prediction of masked words where different context encodings are used for both types of objectives. The proposed approach is evaluated on COCO Captions, Flickr 30K Captions and VQA 2.0 and obtains encouraging results compared to previous methods on image captioning and VQA [199].

Universal image-text representation (UNITER) [43] performs pre-training on four large-scale visual-linguistic datasets (MS-COCO [75], Visual Genome [200], Conceptual Captions [196] and SBU Captions [201]). The learned representations transfer well on downstream tasks such as VQA, Multi-modal retrieval, Visual Commonsense reasoning, and NLVR. In order to emphasize on learning the relationships between visual and language domains, [43] specifically designs pre-training tasks to predict masked visual or text region conditioned on the other domain input, and align language and visual inputs on both the global (image-text) and local (word-region) levels (Fig. 12-a). These tasks are beside the conventional masked language modeling task used in BERT and explicitly include fine-grained word-region alignment alongside conditional masking of inputs that were not considered in the earlier works such as VL-BERT [22], Visual-BERT [63], Vilbert [181] and Unicoder-VL [180]. Common to the other approaches, they adopt the Transformer architecture proposed in BERT that operates on both the visual and language embeddings. In contrast to applying independent Transformers to the language and visual inputs (as in ViLBERT [181] and LXMR [21]),

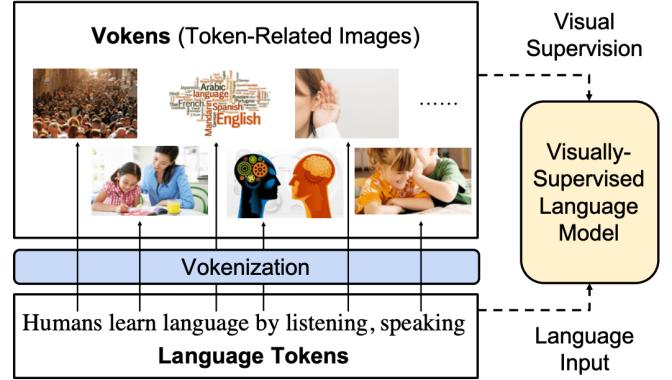


Fig. 13: Visualized tokens (Vokens) [191]: A language model is visually supervised using closely related images that leads to better feature representations from the pretrained model. Figure from [191].

UNITER adopts a single Transformer applied to the textual and image inputs like [22], [63], [180].

VisualBert [63], Uniter [43], VL-BERT [22], VilBERT [181], and Unicoder-VL [180] models for VLP concatenate image and text features and leave it to the self-attention to automatically discover cross-modal relationships. This can complicate the visual grounding of semantic concepts in an image. To address this problem, Object-Semantics Aligned Pre-Training (Oscar) [44] first uses an object detector to obtain object tags (labels), which are then subsequently used as a mechanism to align relevant visual features with the semantic information (Fig. 12-b). The motivation is that the textual content generally pertains to major objects in the image, therefore by explicitly adding those image labels to the input, visual features can be better attended. Similar to BERT [3], Oscar uses a Masked Token Loss for VLP, where different tokens in the textual input and image tags are randomly masked and the model predicts these missing tokens. Further, it also uses a contrastive loss that discriminates between the original and noisy/fake image-tag pairs. The representations thus learned are fine-tuned on VQA, cross-modality retrieval, natural language reasoning, and image captioning tasks to obtain better performances compared to VLP methods that do not use object tags. The recent VinVL [202] approach extends Oscar for the object detection task and learns object instance-centered relationships between visual and language domains using an adapted pretraining scheme. The model is trained on a collection of datasets (MS-COCO, OpenImages, Visual Genome and Objects365) and was demonstrated to precisely relate semantic attributes with the visual information and provided better transferability to the downstream visual comprehension tasks.

3.7.3 Transformers for Visual Grounding

Modulated DETR (MDETR) [203] has a CNN and BERT backbone to extract features from image and text inputs, respectively. The visual and text features are then separately linearly projected to a shared space, concatenated and fed to a transformer model (with an architecture similar to DETR) to predict the bounding boxes for objects corresponding to the queries in the grounding text. The model is trained by

using a loss which predicts a uniform distribution over all relevant text query tokens specific to the predicted bounding boxes. An additional contrastive loss term ensures correspondence between visual and text embedding. TransVG [204] is a simple design, where visual and text features are fused together in a transformer module, and the bounding box corresponding to the query is directly regressed using a learnable token (input to the Transformer module, along-with visual and text features). Referring Transformer [205] is also a simple one stage design where the text and image features are fused in a Transformer encoder, and the Transformer based decoder then directly regresses bounding boxes or segmentation masks. Visual Grounding with Transformer [206] has an encoder-decoder architecture, where visual tokens (features extracted from a pretrained CNN model) and text tokens (parsed through an RNN module) are processed in parallel with two distinct branches in the encoder, with cross-modality attention to generate text-guided visual features. The decoder then computes attention between the text queries and visual features and predicts query-specific bounding boxes.

3.8 Video Understanding

Existing approaches for audio-video data analysis generally learn representations on short-length videos (up to a few seconds long), that allow them to encode only short-range dependencies [1], [32]. Long-range dependency modeling is desirable in various uni-modal and multi-modal learning tasks such as activity recognition [71], [187], [207]–[209]. Below, we explain recent approaches that seek to resolve this challenge using the expressivity of Transformer networks. It is important to note that several of these works [17], [18], [182], [210] still employ (pretrained) CNNs to encode image/frame-level features in the videos on top of which Transformers are applied to model wide context. A few exceptions include [209], [211]–[213] which obtain frame-level features also using the ViT based backbones.

3.8.1 Joint Video and Language Modeling

The VideoBERT [17] model leverages Transformer networks and the strength of self-supervised learning to learn effective multi-modal representations. Specifically, VideoBERT uses the prediction of masked visual and linguistic tokens as a pretext task (Fig. 12-c). This allows modeling high-level semantics and long-range temporal dependencies, important for video understanding tasks. Given a video, [17] converts speech to text using off-the-shelf speech recognition systems and applies vector quantization (clustering) to obtain visual features from pre-trained video classification models. The BERT model is then directly applied to these concatenated sequences of language and visual tokens to learn their joint distribution. The model can be trained with only-text, video-only, and video+text domains. The resulting model showcases interesting capabilities for cross-modal predictions such as video generation from a given textual input (*e.g.*, captions or cooking recipe) and (video-based) future forecasting. The video+text model uses a visual-linguistic alignment task to learn cross-modality relationships. The definition of this pre-text task is simple, given the latent state of the [cls] token, the task is to predict whether the

sentence is temporally aligned with the sequence of visual tokens. Further, the learned representations are shown to be very useful for downstream tasks such as action classification, zero-shot classification, and video captioning.

Zhou *et al.* [210] explore Masked Transformers for dense video captioning. This requires generating language descriptions for all events occurring in a video. Existing works on this problem generally operate sequentially *i.e.*, first detect events and then generate captions in separate sub-blocks. [210] proposes a unified Transformer network to tackle both tasks jointly, thereby seamlessly integrating the multi-modal tasks of event detection and captioning. First, a video encoder is used to obtain frame-wise representations followed by two decoder blocks focused on proposing the video events and the captions. Since untrimmed videos are considered, a masking network is used in the captioning decoder to focus on describing a single event proposal. Remarkably, [210] was the first approach to target dense video captioning using non-recurrent models and used self-attention in the encoder(applied on CNN derived features) to model broad range context between video frames. Experiments on ActivityNet Captions [214] and YouCookII [215] datasets showed good improvements over previous recurrent network and two-stage based approaches.

3.8.2 Video Action Recognition

The traditional CNN based methods in video classification generally perform 3D spatio-temporal processing over limited intervals to understand videos. Neimark *et al.* [211] propose Video Transformer Network (VTN) that first obtains frame-wise features using 2D CNN and apply a Transformer encoder (Longformer [103]) on top to learn temporal relationships. Longformer is an attractive choice to process long sequences (with an arbitrary length n) due to its $\mathcal{O}(n)$ complexity. The classification token is passed through a fully connected layer to recognize actions or events. The advantage of using Transformer encoder on top of spatial features is two fold: (a) it allows processing a complete video in a single pass, and (b) considerably improves training and inference efficiency by avoiding the expensive 3D convolutions. This makes VTN particularly suitable for modeling long videos where interactions between entities are spread throughout the video length. Their experiments on Kinetics-400 dataset [71] with various backbones (ResNet [67], ViT [11] and DeiT [12]) shows competitive performance.

Girdhar *et al.* [18] use a variant of Transformer architecture to aggregate person-specific contextual cues in a video for action classification and localization. Initially, the model uses a Faster-RCNN [125] style processing where a backbone model generates features that are forwarded to the Region Proposal Network to obtain object proposals. Then RoI pooling is applied to generate object-specific features. Multi-head self-attention [1] is then applied on top of the object features as a cascade of self-attention layers. In each Transformer unit, a particular person feature is treated as the ‘query’ (Q), while the features from the neighboring video clip are used as ‘key’ (K) and ‘value’ (V). The location information is explicitly encoded in the input feature map from which K, V and Q are derived, thus incorporating the positional information in the self-attention. For a given $400 \times 400 \times 64$ video clip, the key and value tensors are of size

$16 \times 25 \times 25 \times 128$, while the query is 128 dimensional vector. Although [18] uses only RGB stream, additional modalities like optical flow and audio signal (as in competing works) would further increase the compute complexity. Further, the Transformer model was found to be sub-optimal for action localization, perhaps due to its tendency to incorporate global information. Therefore, it is important to achieve the right trade-off between the global and local context for problems that demand precise delineation (*e.g.*, action localization and segmentation).

Human action recognition based on skeleton representation requires understanding relationships between different joints of a body in a given frame as well as between different frames of a video. Plizzari *et al.* [216] proposed a two-stream Transformer network to model such relationships. They introduced spatial self-attention (SSA) to model relations between different body-joints (Fig. 14a) while temporal self-attention (TSA) to capture long-range inter-frame dependencies (Fig. 14b). They first used a small residual network to extract features from skeleton data and then used SSA and TSA modules to process those feature maps. SSA finds the correlation between each pair of joints independently, while TSA focuses on how features of a certain joint change between frames along the temporal dimension. The purpose of SSA is to discover relationships among the surrounding joints in the same way as the Transformer relates different words in a phrase. On the other hand, TSA finds long-range relations between frames, similar to how relations among phrases are built in NLP. The two streamed model achieves state-of-the-art results on NTU-RGB+D 60 [217] and NTU-RGB+D 120 [218] datasets.

Multiscale Vision Transformers (MViT) [219] build a feature hierarchy by progressively expanding the channel capacity and reducing the spatio-temporal resolution in videos. They introduce multi-head pooling attention to gradually change the visual resolution in their pyramid structure. TimeSFormer [213] extends ViTs [132] to videos, by considering the video as a sequence of patches extracted from individual frames. To capture spatio-temporal relationships, they propose divided attention *i.e.*, spatial and temporal attentions are separately applied within each block. TimeSFormer demonstrates SOTA performance on action recognition, and can be applied to clips over one minute. Another notable pure-transformer based model is the Video Vision Transformer (ViViT) [212]. First, the spatio-temporal tokens are extracted and then efficient factorised versions of self-attention are applied to encode relationships between tokens. However, they require initialization with image-pretrained models to effectively learn the ViT models. There has also been concurrent work on learning sound pretrained models using self-supervised learning with ViTs. An important recent effort is the long-short contrastive learning (LSTCL) framework [220], which reconstructs representations from different time-scales (narrow and broad) as auxiliary learning tasks and demonstrates good downstream performance.

3.8.3 Video Instance Segmentation

The Video Instance Segmentation Transformer (VisTR) [209] model extends DETR [13] for video object instance segmentation (VIS) task. Local features are obtained using a

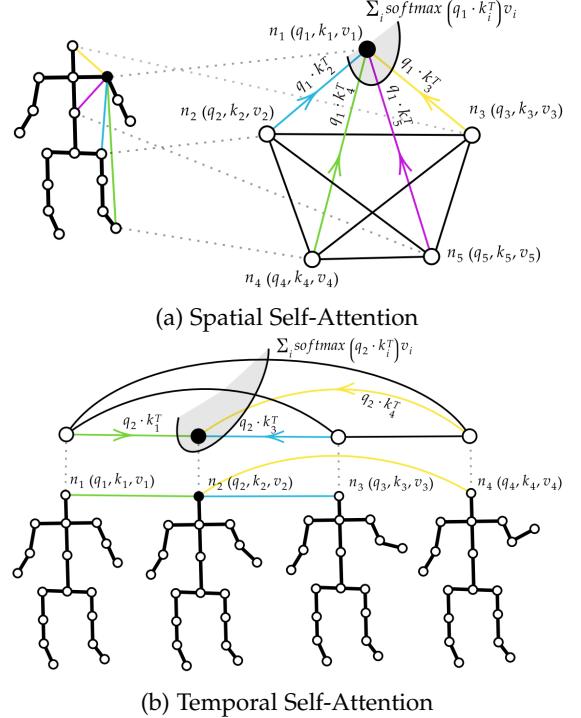


Fig. 14: Spatial/Temporal Attention for Skeleton Data Representations. Relationships between body-joints and inter-frame dependencies are modeled using two dedicated self-attention modules. Figure is from [216].

backbone CNN on a collection of video frames. An encoder and a decoder Transformer is used similar to DETR to frame the instance segmentation problem as a sequence to sequence prediction task. The input frame-level features are concatenated to form clip representations and the Transformer outputs instance predictions in a order that is consistent across frames. This integrates the object detection and tracking with-in a single unified architecture. The predicted outputs are matched with the ground-truth using bipartite matching. Similar to Mask R-CNN [127], a separate head is used to predict the instance mask based on self-attention and 3D convolutions. The overall results are competitive among the single model approaches on YouTube VIS dataset [221], but performs somewhat lower compared to more complex CNN-based models such as MaskProp [222].

3.9 Transformers in Low-shot Learning

In the few-shot learning settings, a support set is provided at the inference to adapt to a novel set of categories. Transformer models have been used to learn set-to-set mappings on this support set [26] or learn the spatial relationships between a given input query and support set samples [25]. In terms of absolute performance, the patch-wise spatial self-attention between query and support set images excels compared to an image level association learned in [26]. However, the patch-wise attention computation is computationally expensive. We elaborate on these approaches below.

Doersch *et al.* [25] explore the utility of self-supervision and Transformer model for few-shot fine-grained classification, where distribution mismatch exists between training

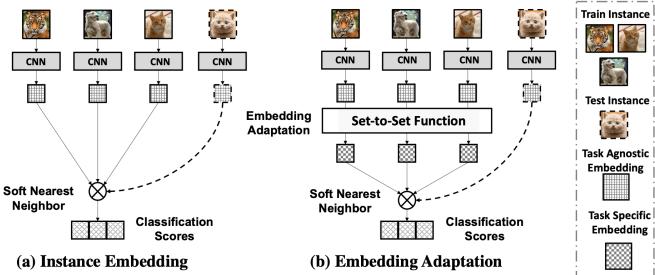


Fig. 15: An overview of FEAT [26]. Compared to the conventional instance embedding methods in FSL that keep the embedding function same for all tasks (a), FEAT uses a set-to-set function to adapt the embedding function to each FSL task (b). It evaluates several set-to-set functions and found the Transformer module to be the most suitable choice for FSL. Figure from [26].

and evaluation phases. They develop Cross-Transformer model to relate a given query image with the few-examples available in the support set. To this end, the Transformer finds spatially similar regions in the query and support set images, and the corresponding features are then used to obtain class decisions for the query. The queries in the Transformer architecture are derived from the grid features obtained using the query image. Similarly, grid features from the support images are used to construct keys and values which are in turn used to derive attended outputs. This approach, besides a contrastive self-supervision based training mechanism, leads to the best performance on the challenging Meta-dataset [223].

Ye *et al.* [26] propose to adapt the few-shot embeddings learned on the base classes to the few-shot target classes during inference using a Transformer module. This leads to task-specific embeddings that perform better on the discriminative tasks such as few-shot classification. While many other set-to-set functions are also evaluated, such as Graph convolutional networks [224], Bidirectional LSTMs [32] and DeepSets [225], the best performance is achieved with the Transformer-based mapping. This is attributed to the better contextualization, task interpolation and extrapolation capability of Transformers and their permutation invariance while maintaining a relatively lower parameter complexity. The Transformer architecture in [26] follows the standard model [1]. The embeddings are adapted using a contrastive loss function for preserving discriminative properties (Fig. 15). The resulting model achieves strong performance on inductive, transductive, and generalized FSL tasks.

Liu *et al.* [226] learn a multi-head self-attention based module, to integrate the visual representation learned by the models trained on different domains present in the meta-dataset [223]. The Universal Representation Transformer (URT) layer dynamically re-weights the representations from different domain-specific backbones, and proves very effective in handling few shot tasks across a variety of data distributions.

3.10 Transformers for Clustering

Clustering aims to discover structure in the data by grouping similar data points together. It has numerous applications such as data visualization and interpretation, anomaly detection, and open-set categorization. Neural networks have been developed for set prediction problems [225], [227], however, the setpoints are processed individually which can lose information about inter-point relationships. Recent works employ Transformers that operate on set inputs called the Set Transformers (ST) [228] for *amortized* clustering. Amortized clustering is a challenging problem that seeks to learn a parametric function that can map an input set of points to their corresponding cluster centers. Lee *et al.* [228] propose to learn such a mapping function using a Transformer architecture comprising of multi-head self-attention blocks [1]. The Transformer model is permutation invariant by design and allows encoding both pair-wise and higher-order relationships between the input points. However, a full Transformer would lead to a high computational cost of $\mathcal{O}(n^2)$ in each self-attention layer, where n is the number of points in the set. ST reduces this cost to $\mathcal{O}(mn)$ by using an Induced Self-Attention Block that uses a low-rank projection ($H \in \mathbb{R}^m$) to allow operating on large sets. The model was trained to learn optimal parameters that maximize the likelihood of a mixture of Gaussians (MoGs). Thus MoG parameters are estimated by the ST given a set of data points. Beyond amortized clustering, ST is a generic framework which can handle other set-input problems such as counting unique elements in an input set, multi-instance learning, set anomaly detection, and 3D point-cloud classification. More recently, [229] improves [228] by taking a sequential approach to cluster generation, thereby allowing assignment to a variable number of clusters.

3.11 Transformers for 3D Analysis

Given the irregular (variable number of points) and permutation invariant nature of 3D point cloud representations, Transformers provide a promising mechanism to encode rich relationships between 3D data points. To this end, recent works [230], [231] are motivated by the capability of Transformers to learn set-functions. Specifically, [230] introduced a Point Transformer which uses vector attention to learn weights for each channel, while [231] suggest an alternate design where local 3D structure is explicitly encoded. The non-local nature of Transformers is exploited in [45] towards an accurate human pose and mesh reconstruction algorithm. We discuss these approaches below.

Self-attention being a set-operator is ideally suited for processing point clouds, a 3D data representation that demands invariance to number of points and their permutations. Zhao *et al.* [230] propose a point Transformer layer that applies self-attention in the local neighborhood of 3D points. The proposed layer builds on vectorized self-attention network (SAN) [82] where attention weights are represented with vectors. Furthermore, a positional encoding is added both to the attention vector and transformed features (value vectors) to represent location information. The point Transformer layer is sandwiched between two linear layers to create a point Transformer block that is stacked multiple times in the developed network architecture. Their design

also included transition down/up blocks to reduce/increase the number of points in the input (in a typical encoding-decoding pipeline style). The resulting architecture shows promising results on the 3D classification and segmentation tasks.

The Point Cloud Transformer (PCT) [231] is a parallel work to [230] and motivated by the permutation invariance property of Transformers. However, compared to [230], it is more directly based on the conventional Transformer architecture [1] and does not involve vector attention. The key modifications include a 3D coordinate-based position encoding, an offset attention module, and a neighbor embedding that encodes local 3D structure in point-clouds. Specifically, the offset attention layer calculates the difference between the self-attended features and the input features using element-wise subtraction. The local neighbor embedding simply finds self-attention relationships among a group of points instead of individual 3D points. Explicitly incorporating local neighbourhood information makes this a more efficient architecture compared to [230]. The method shows promising performance on 3D shape classification, normal estimation and segmentation tasks on ModelNet40 [232] and ShapeNet [233] datasets.

The Mesh Transformer (METRO) [45] model targets 3D human pose and mesh reconstruction from a single 2D image. A key challenge here is to faithfully learn the non-local interactions between body-joints and mesh vertices (*e.g.*, hand and foot). The expressivity of Transformer network is used to jointly model *vertex to vertex* relationships in a mesh as well as the *vertex to body-joint* relationships. The self-attention mechanism can attend to any combination of vertices in the mesh, thereby encoding non-local relationships. The multi-layer Transformer architecture sequentially performs dimensionality reduction to map the 2D image to 3D mesh. Position encoding is performed using the 3D coordinates (x, y, z) of each vertex and each body-joint. Similar to masked language modeling in NLP, METRO uses masked vertex modeling (MVM) which randomly masks some percentage of input queries (see Fig. 16). The Transformer is tasked with regressing all the joints and vertices which helps encode inter-dependencies between them. METRO obtains state-of-the-art results on human mesh reconstruction on Human3.6M [234] and 3DPW [235] datasets. Since the approach does not depend on a parametric mesh model, it generalizes well to other reconstruction tasks such as 3D hand reconstruction [236]. Overall, this is the first effort to employ Transformers for 3D human reconstruction tasks and leads to fairly good results.

4 OPEN CHALLENGES & FUTURE DIRECTIONS

Despite excellent performance from Transformer models and their interesting salient features (Table 1), there exist several challenges associated with their applicability to practical settings (Table 2). The most important bottlenecks include requirement for large-amounts of training data and associated high computational costs. There have also been some challenges to visualize and interpret Transformer models. In this section, we provide an overview of these challenges, mention some of the recent efforts to address those limitations and highlight the open research questions.

4.1 High Computational Cost

As discussed in Sec. 1, a strength of Transformer models is their flexibility to scale to high parametric complexity. While this is a remarkable property that allows training enormous sized models, this results in high training and inference cost (a detailed comparison between CNN and ViTs is shown in Table 3). As an example, the BERT [3] basic model (with 109 million parameters) took around 1.89 peta-flop days² for training, while the latest GPT3 [6] model (175 billion parameters) took around 3640 peta-flop days for training (a staggering $\sim 1925 \times$ increase). This comes with a huge price tag, *e.g.*, according to one estimate [237], GPT3 training might have cost OpenAI 4.6 million USD. Additionally, these large-scale models require aggressive compression (*e.g.*, distillation) to make them feasible for real-world settings.

An empirical study on the scalability of Vision Transformers for number of parameters (ranging from five million to two billion), size of the training datasets (ranging from 30 million to three billion training images), and compute budget (1-10000 TPU core-days) is presented in [238]. From this study, We can draw the following conclusions (a) scaling up on compute, model and size of training samples improves performance (b) only large models (with more parameters) can benefit from more training data, and the performance of smaller models plateau quickly and can not leverage from additional data. This indicates that large scale models have the capacity to further enhance their representation learning capabilities. However, with the current designs, scaling upon Transformer models is expensive and compute prohibitive, thus necessitating the need for efficient designs.

In the language domain, recent works focus on reducing the high complexity of Transformer models (basically arising from the self-attention mechanism [1] where a token's representation is updated by considering all tokens from the previous layer). For example, [103], [245] explore selective or sparse attention to previous layer tokens while updating each next layer token. Linformer [38] reduces complexity of standard self-attention operation from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ (both in time and memory requirements). The main idea is to show that a low-rank matrix is sufficient to model the self-attention mechanism. The Reformer model [246] employed locally-sensitive hashing (LSH) to minimize the complexity of self-attention from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log(n))$. In similar pursuit, the recent Lambda Networks propose to model local context as a linear function which helps reduce complexity of self-attention [247]. These linear function lambdas are applied to the input query to model contextual relationships between pixels.

Vyas *et al.* [248] developed an efficient *cluster attention* to deal with large input sequences that approximates the original self-attention. The cluster attention groups queries into clusters and then computes attention between cluster centers (instead of attention between all the queries that leads to quadratic complexity). The main idea is that the queries close in the Euclidean space should have similar attention distributions. With a fixed number of clusters, this intuition helps reduce the quadratic complexity to linear

2. A peta-flop day is a measure of computation and equals to performing 10^{15} neural net operations per second for one complete day.

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
Image Classification	ViT [11]	Directly adopted NLP Transformer Encoder for images, Mechanism to linearly embed image patches with positional embedding suitable for the Encoder.	2D Image	Class labels	Cross-entropy
	DeiT [12]	Transformer as student while CNN as a teacher, Distillation tokens to produce estimated labels from teacher, Attention between class and distillation tokens.	2D Image	Class labels	Cross-entropy, Distillation loss based on KL-divergence
	CLIP [195]	Jointly train image and text encoders on image-text pairs, to maximize similarity of valid pairs and minimize otherwise	2D Images & texts	Image-text pairs	Symmetric cross-entropy
Object Detection	DETR [13]	Linear projection layer to reduce CNN feature dimension, Spatial positional embedding added to each multi-head self-attention layer of both encoder and decoder. Object queries (output positional encoding) added to each multi-head self-attention layer of decoder.	2D Image	Class labels	Hungarian loss based on bipartite matching between predicted and ground truths
	D-DETR [14]	Deformable Transformer consists of deformable attention layers to introduce sparse priors in Transformers, Multi-scale attention module.	2D Image	Class labels	Hungarian loss
Low Shot Learning	CT [25]	Self-supervised pretraining, Query-aligned class prototypes that provide spatial correspondence between the support-set images and query image.	2D Image	Pretraining without labels and few-shot learning with Class labels	Normalized Cross-entropy
Image Colorization	CoTran [24]	Conditional Row/column multi-head attention layers, Progressive multi-scale colorization scheme.	2D Image	2D Image	Negative log-likelihood of the images
Action Recognition	ST-TR [216]	Spatial and Temporal self-attention to operates on graph data such as joints in skeletons.	Skeleton	Action Classes	Cross-entropy
Super-resolution	TTSR [16]	Texture enhancing Transformer module, Relevance embeddings to compute the relevance between the low-resolution and reference image.	2D Image	2D Image	Reconstruction loss, Perceptual loss defined on pretrained VGG19 features.
Multi-Model Learning	Oscar [44]	Transformer layer to jointly process triplet representation of image-text [words, tags, features], Masked tokens to represent text data.	2D Image	Captions, Class labels, Object tags	Negative log-likelihood of masked tokens, Contrastive binary cross-entropy
3D Classification/Segmentation	PT [230]	Point Transformer block, Transition down block to reduce cardinality of the point set, Transition up for dense prediction tasks.	CAD models, 3D object part segmentation	Object and shape categories	Cross-entropy
3D Mesh Reconstruction	METRO [45]	Progressive dimensionality reduction across Transformer layers, Positional Encoding with 3D joint and 3D vertex coordinates, Masked vertex/joint modeling.	2D Image	3D Mesh + Human Pose	L_1 loss on mesh vertices and joints in 3D and 2D projection.
Vision and Language Navigation	Chen <i>et al.</i> [194]	Uni-modal encoders on language and map inputs followed by a cross-modal transformer, Trajectory position encodings in the map encoder.	Instruction text + RGBD panorama + Topological Environment Map	Navigation Plan	Cross-entropy over nodes and [stop] action
Referring Image Segmentation	CMSA [15]	Multimodal feature, Cross-modal self-attention on multiple levels and their fusion using learned gates.	2D Image + Language expression	Segmentation mask	Binary cross-entropy loss
Video Classification	Lee <i>et al.</i> [182]	Operates on real-valued audio-visual signals instead of tokens, Contrastive learning for pre-training, End-to-end multi-modal transformer learning.	Audio-Visual	Activity labels	Contrastive Infonce loss and Binary cross-entropy

TABLE 1: A summary of key design choices adopted in different variants of transformers for a representative set of computer vision applications. The main changes relate to specific loss function choices, architectural modifications, different position embeddings and variations in input data modalities.

Task	Method	Metric	Dataset	Performance	Highlights	Limitations
Image Classification	ViT [1] ICLR'21	Top-1 Acc.	ImageNet	88.55	a) First application of Transformer (global self-attention) directly on image patches, b) Convolution-free network architecture, c) Outperforms CNN models such as ResNet.	a) Requires training on large-scale data e.g., 300-Million images, b) Requires careful transfer learning to the new task, c) Requires large model with 632-Million parameters to achieve SOTA results.
	DeiT [12] arXiv'20	Top-1 Acc.	ImageNet	83.10	a) Successfully trains Transformer on ImageNet only, b) Introduces attention-based distillation method. c) Produces competitive performance with small (86-Million parameters) Transformers.	a) Requires access to pretrained CNN based teacher model thus performance depends on the quality of the teacher model.
Low-Shot Learning	Swin-T [36] arXiv'21	Top-1 Acc.	ImageNet	84.5	a) Provides a general purpose backbone for different vision tasks e.g., classification, detection and segmentation b) A hierarchical design using shifted-windows operation.	a) Hard to train from scratch on smaller datasets b) Quadratic compute complexity inherent to the self-attention operation.
	CT [25] NeurIPS'20	Top-1 Acc.	ImageNet COCO	62.25 60.35	a) Self-supervised pre-training mechanism that does not need manual labels, b) Dynamic inference using Transformer achieving stat-of-the-art results.	Proposed algorithm is limited in its capacity to perform on datasets that lack spatial details such as texture.
Object Detection	DETR [13] ECCV'20	AP	COCO	44.9	a) Use of Transformer allows end-to-end training pipeline for object detection, b) Removes the need for hand-crafted post-processing steps.	a) Performs poorly on small objects, b) Requires long training time to converge.
	D-DETR [14] ICLR'21	AP	COCO	43.8	a) Achieves better performance on small objects than DETR [13], b) Faster convergence than DETR [13]	Obtain SOTA results with 52.3 AP but with two stage detector design and test time augmentations.
Image Colorization	ColTran [24] ICLR'21	FID	ImageNet	19.71	a) First successful application of Transformer to image colorization, b) Achieves SOTA FID score.	a) Lacks end-to-end training, b) limited to images of size 256×256.
Action Recognition	ST-TR [216] arXiv'20	Top-1 Acc.	NTU 60/120	94.0/84.7	a) Successfully applies Transformer to model relations between body joints both in spatial and temporal domain, b) Achieves SOTA results.	Proposed Transformers do not process joints directly rather operate on features extracted by a CNN, thus the overall model is based on hand-crafted design.
Super-Resolution	TTSR [16] CVPR'20	PSNR/ SSIM	CUFED5 Sun80 Urban100 Manga109	27.1 / 0.8 30.0 / 0.81 25.9 / 0.78 30.1 / 0.91	a) Achieves state-of-the-art super-resolution by using attention, b) Novel Transformer inspired architectures that can process multi-scale features.	a) Proposed Transformer does not process images directly but features extracted by a convolution based network, b) Model with large number of trainable parameters, and c) Compute intensive.
Multi-Model Learning	ViLBERT [181] NeurIPS'19	Acc./ mAP ($R@1$)	VQA [183]/ Retrieval [239]	70.6 / 58.2	a) Proposed Transformer architecture can combine text and visual information to understand inter-task dependencies, b) Achieves pre-training on unlabelled dataset.	a) Requires large amount of data for pre-training, b) Requires fine tuning to the new task.
Oscar [44] ECCV'20	Acc./ mAP ($R@1$)	VQA [240]/ COCO	80.37/57.5	a) Exploit novel supervisory signal via object tags to achieve text and image alignment, b) Achieves state-of-the-art results.	Requires extra supervision through pre-trained object detectors thus performance is dependent on the quality of object detectors.	
UNITER [43] ECCV'20	Acc./ Avg. ($R@1/5/10$)	VQA [183]/ Flickr30K [241]	72.47/83.72	Learns fine-grained relation alignment between text and images	Requires large multi-task datasets for Transformer training which lead to high computational cost.	
3D Analysis	Point Transformer [230] arXiv'20	Top-1 Acc. IoU	ModelNet4092.8 [232]	85.9	a) Transformer based attention capable to process unordered and unstructured point sets, b) Permutation invariant architecture.	a) Only moderate improvements over previous SOTA, b) Large number of trainable parameters around 6× higher than PointNet++ [242].
METRO [45] arXiv'20	MPJPE PA-MPJPE MPVE	3DPW [235]	77.1 47.9 88.2	a) Does not depend on parametric mesh models so easily extendable to different objects, b) Achieves SOTA results using Transformers.	Dependent on hand-crafted network design.	

TABLE 2: A summary of advantages and limitations of different Transformers based methods in different Tasks. (CT: Cross Transformers, AP: Average Precision, mAP: mean AP, IoU: Intersection over Union, FID: Fréchet inception distance, MPJPE: Mean Per Joint Position Error, MPVE: Mean Per Vertex Error).

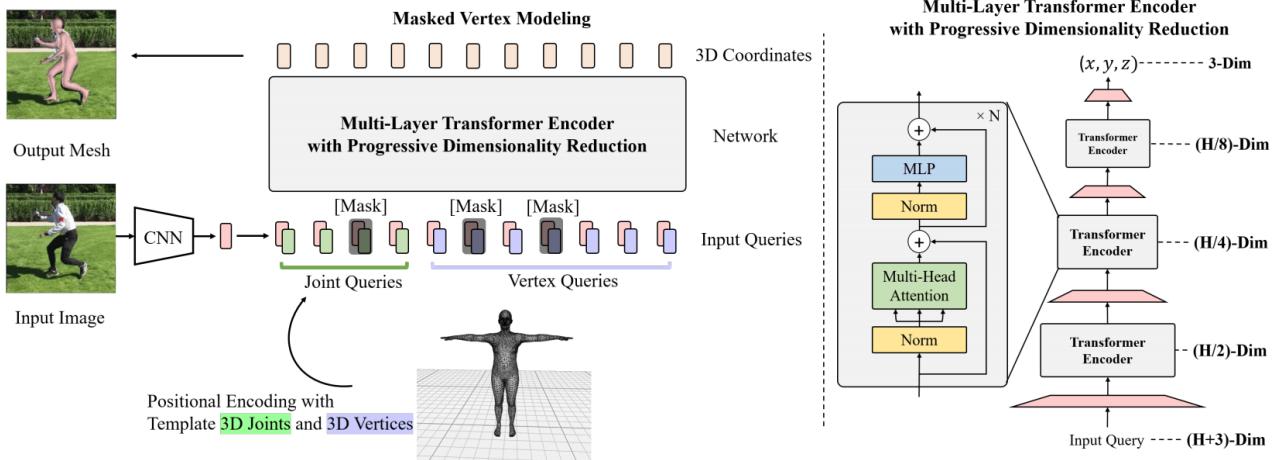


Fig. 16: Mesh Transformer architecture. The joint and vertex queries are appended with positional embeddings and passed through multiple self-attention layers to jointly regress 3D coordinates of joints and mesh vertices. Figure is from [45].

Method	#Param (M)	GFLOPs	Top-1 Acc (%)	Method	#Param (M)	GFLOPs	Top-1 Acc (%)
ResNet18 [67]*	11.7	1.8	69.8	ResNet101 [67] *	44.7	7.9	77.4
EfficientNet-B3 [87]*	12.0	1.8	81.6	ResNeXt101-32x4d [244]*	44.2	8.0	78.8
DeiT-T [12]	5.7	1.3	72.2	RegNetY-8G [86]*	39.0	8.0	81.7
T2T-ViT _t -7 [35]	5.0	1.3	71.7	EfficientNet-B5 [87] *	30.0	9.9	83.6
LocalViT-T [107]	5.9	1.3	74.8	CvT-21 [96]	32.0	7.1	82.5
CrossViT-T [104]	6.9	1.6	73.4	CaiT-S-24 [243]	32.2	9.4	82.7
PVTv1-T [93]	13.2	1.9	75.1	T2T-ViT _t -19 [35]	39.0	9.8	81.4
ResT-Lite [110]	10.5	1.4	77.2	PVTv1-M [93]	44.2	6.7	81.2
CaiT-XXX-24 [243]	12.0	2.5	77.6	PVTv2-B3 [97]	45.2	6.9	83.2
PVTv2-B1 [97]	13.1	2.1	78.7	NesT-S [111]	38.0	10.4	83.3
Lv-ViT-T [89]	8.5	—	79.1	ResNet152 [67] *	60.2	11.6	78.3
RegionViT-T [100]	13.8	2.4	80.4	CaiT-S-36 [243]	48.0	13.9	83.3
ResNet50 [67]*	25.6	4.1	76.1	T2T-ViT _t -24 [35]	64.0	15.0	82.2
ResNeXt50-32x4d [244]*	25.0	4.3	77.6	PVTv1-L [93]	61.4	9.8	81.7
RegNetY-4G [86]*	21.0	4.0	80.0	TNT-B [88]	66.0	14.1	82.8
EfficientNet-B4 [87]*	19.0	4.2	82.9	Swin-S [36]	50.0	8.7	83.0
DeiT-S [12]	22.1	4.6	79.9	Twins-SVT-B [37]	56.0	8.3	83.2
PVTv1-S [93]	24.5	3.8	79.8	RegionViT-B [100]	72.7	13.0	83.3
LocalViT-S [107]	22.4	4.6	80.8	PVTv2-B4 [97]	62.6	10.1	83.6
CrossViT-S [104]	26.7	5.6	81.0	ResNeXt101-64x4d [244] *	83.5	15.6	79.6
TNT-S [88]	23.8	5.2	81.3	RegNetY-16G [86] *	84.0	16.0	82.9
Swin-T [36]	29.0	4.5	81.3	EfficientNet-B6 [87] *	43.0	19.0	84.0
NesT-T [111]	17.0	5.8	81.5	NesT-B [111]	68.0	17.9	83.8
T2T-ViT _t -14 [35]	21.5	5.2	81.5	ViT-B/16 [11]	86.6	17.6	79.8
CvT-13 [96]	20.0	4.5	81.6	DeiT-B/16 [12]	86.6	17.6	81.8
ResT-B [110]	30.3	4.3	81.6	Swin-B [36]	88.0	15.4	83.3
Twins-SVT-S [37]	24.0	2.8	81.7	Twins-SVT-L [37]	99.2	14.8	83.7
PVTv2-B2-Li [97]	22.6	3.9	82.1	PVTv2-B5 [97]	82.0	11.8	83.8
RegionViT-S [100]	30.6	5.6	82.5	Lv-ViT-M [89]	56.0	16.0	84.1
Lv-ViT-S [89]	26.0	6.6	83.3				

TABLE 3: A Comparative analysis between different vision transformer and CNN models in terms of their parameter complexity and top-1 (%) accuracy on ImageNet validation set. For a direct comparison, we consider models that are trained on ImageNet from scratch on input of size 224x224. * denotes pure CNN-based methods.

complexity of $\mathcal{O}(nc)$ with respect to the input sequence length n (where c is the number of clusters). We refer interested readers to a survey on efficient Transformers in NLP [34].

Similar to the NLP domain, computer vision models also suffer from the high computational cost of Transformer models. For example, image generators that are based on sequence-based Transformers (e.g., iGPT) have a high compute cost limiting their applicability to high-resolution inputs. The time and memory cost of core self-attention operation in Transformers increases quadratically with the number of patches, i.e. $\mathcal{O}(n^2)$, for n image patches (in some

applications, e.g., low-level vision, $n = H \times W$ where H, W denote the height and width of the image). This is a major drawback of existing Transformers that hinders their application to most tasks involving high-resolution (HR) images, such as object detection and segmentation (in high-level vision), and super-resolution, deblurring, denoising, etc. (in low-level vision). Numerous methods have been proposed that make special design choices to perform self-attention more ‘efficiently’, for instance employing pooling/downsampling in self-attention [97], [219], [249], local window-based attention [36], [250], axial-attention [179], [251], low-rank projection attention [38], [252], [253], ker-

nelizable attention [254], [255], and similarity-clustering based methods [246], [256]. However, almost all of these approaches either come with a trade-off between complexity and accuracy, require special hardware specifications or are still not applicable to very large images. Therefore, there is a pressing need to develop an efficient self-attention mechanism that can be applied to HR images on resource-limited systems without compromising accuracy. It will be interesting to explore how existing models can be extended to high-dimensional cases *e.g.*, using a *multi-scale transformer* design with a somewhat local context modeling. By inducing inductive biases based on our understanding of the visual learning tasks (e.g., spatial relationships in the local neighbourhood), the high computational cost can be reduced. Similarly, using sparse attention maps modeled with low-rank factorization in the matrices can also help towards reducing the computational cost [211].

4.2 Large Data Requirements

Transformer architectures do not inherently encode inductive biases (prior knowledge) to deal with visual data, they typically require large amount of training to figure out the underlying modality-specific rules. For example, a CNN has inbuilt translation invariance, weight sharing, and partial scale invariance due to pooling operations or multi-scale processing blocks. However, a Transformer network needs to figure out these image-specific concepts on its own from the training examples. Similarly, relationships between video frames need to be discovered automatically by the self-attention mechanism by looking at a large database of video sequences. This results in longer training times, a significant increase in computational requirements, and large datasets for processing. For example, the ViT [11] model requires hundreds of millions of image examples to obtain reasonable performance on the ImageNet benchmark dataset. The question of learning a Transformer in a data-efficient manner is an open research problem and recent works report encouraging steps towards its resolution. For example, DeiT [12] uses a distillation approach to achieve data efficiency while T2T (Tokens-to-Token) ViT [35] models local structure by combining spatially close tokens together, thus leading to competitive performance when trained only on ImageNet from scratch (without pre-training). By incorporating CNNs like feature hierarchies in ViTs to effectively capture local image cues, ViTs (*e.g.*, CCT [106], NesT [111]) can be trained from scratch even on small-scale datasets (*e.g.*, CIFAR-10). Another approach to data efficient training of ViTs is proposed in *et al.* [257]. The authors show that by smoothing the local loss surface using sharpness-aware minimizer (SAM) [258], ViTs can be trained with simple data augmentation scheme (random crop, and horizontal flip) [259], instead of employing compute intensive strong data augmentation strategies, and can outperform their counterpart ResNet models.

4.3 Vision Tailored Transformer Designs

We note that most of the existing works focused on vision tasks tend to directly apply NLP Transformer models on computer vision problems. These include architectures designed for image recognition [11], video understanding [17]

and especially multi-modal processing [181]. Although the initial results from these simple applications are quite encouraging and motivate us to look further into the strengths of self-attention and self-supervised learning, current architectures may still remain better tailored for language problems (with a sequence structure) and need further intuitions to make them more efficient for visual inputs. For example, vector attention from [82] is a nice work in this direction which attempts to specifically tailor self-attention operation for visual inputs via learning channel-wise attentions. Similarly, [260] uses a Jigsaw puzzle based self-supervision loss as a parallel branch in the Transformers to improve person re-identification. A recent work [35] rearranges the spatially close tokens to better model relationships in spatially proximal locations. Token distillation [12] from pre-trained CNN models has also been used as a remedy to inject domain biases in the representations. One may argue that the architectures like Transformer models should remain generic to be directly applicable across domains, we notice that the high computational and time cost for pre-training such models demands novel design strategies to make their training more affordable on vision problems.

4.4 Neural Architecture Search for ViTs

While Neural Architecuter Search (NAS) has been well explored for CNNs to find an optimized architecture, it is relatively less explored in Transformers (even for language transformers [261], [262]). Chen *et al.* [263] propose a one-shot NAS for vision transformers, called AutoFormer. BossNAS [264] searches for a hybrid architecture (CNN and Transformer). Another recent effort studies the trade-off between global and local information in Transformers in the context of vision applications [265]. It will be insightful to further explore the domain-specific design choices (*e.g.*, the contrasting requirements between language and vision domains) using NAS to design more efficient and lightweight models similar to CNNs [87].

4.5 Interpretability of Transformers

Through an extensive set of carefully designed experiments, Naseer *et al.* [266] investigate multiple intriguing properties of ViTs in terms of their generalization and robustness. They show that, compared with CNNs, ViTs demonstrate strong robustness against texture changes and severe occlusions, *e.g.* ViTs retain upto 60% top-1 accuracy on ImageNet once 80% of the image content is randomly occluded. Given the strong performance of Transformer architectures, it is interesting and critical to interpret their decisions, *e.g.*, by visualizing relevant regions in an image for a given classification decision. The main challenge is that the attention originating in each layer, gets inter-mixed in the subsequent layers in a complex manner, making it difficult to visualize the relative contribution of input tokens towards final predictions. This is an open problem, however, some recent works [267]–[269] target enhanced interpretability of Transformers and report encouraging results. Attention roll-out and attention flow methods were proposed in [268] to estimate the accurate attentions. However, this method functions in an ad-hoc manner and makes simplistic assumptions *e.g.*, input tokens are

linearly combined using attention weights across the layers. Chefer *et al.* [269] note that the attention scores obtained directly via the self-attention process (encoding relationships between tokens) or reassessments in [268] do not provide an optimal solution. As an alternative, they propose to assign and propagate *relevancy scores* in the Transformer network such that the sum of relevancy is constant throughout the network. Their design can handle both the positive and negative attributions experienced in the self-attention layer. The proposed framework has an added advantage of being able to provide class-specific visualizations. Despite these seminal works, visualizing and interpreting Transformers is an unsolved problem and methods are needed to obtain spatially precise activation-specific visualizations. Further progress in this direction can help in better understanding the Transformer models, diagnosing any erroneous behaviors and biases in the decision process. It can also help us design novel architectures that can help us avoid any biases.

4.6 Hardware Efficient Designs

Large-scale Transformer networks can have intensive power and computation requirements, hindering their deployment on edge devices and resource-constrained environments such as internet-of-things (IoT) platforms. Some recent efforts have been reported to compress and accelerate NLP models on embedded systems such as FPGAs [270]. Li *et al.* [270] used an enhanced block-circulant matrix-based representation to compress NLP models and proposed a new Field Programmable Gate Array (FPGA) architecture design to efficiently manage resources for high throughput and low latency. They could achieve 27x, 3x and 81x improvements in performance (throughput measured in FPS), reduced power consumption, and energy efficiency relative a CPU for RoBERTa model [7]. Towards this goal, [262] proposed to design Hardware-Aware Transformers (HAT) using neural architecture search strategies [271]–[273]. Specifically, a SuperTransformer model is first trained for performance approximation which can estimate a model’s performance without fully training it. This model comprises the largest possible model in the search space while sharing weights between common parts. Eventually, an evolutionary search is performed considering the hardware latency constraints to find a suitable SubTransformer model for a target hardware platform (*e.g.*, IoT device, GPU, CPU). However, such hardware efficient designs are currently lacking for the vision Transformers to enable their seamless deployment in resource-constrained devices. Further, the search cost of the evolutionary algorithms remains significant with the associated impact of CO₂ emissions on the environment.

4.7 Towards Integrating All Modalities

Since Transformers provide a unified design to process different modalities, recent efforts also focus on proposing more generic general purpose reasoning systems based on Transformers. Inspired by the biological systems that can process information from a diverse range of modalities, Perceiver model [274] aims to learn a unified model that can process any given input modality without making domain-specific architectural assumptions. In order to scale

to high-dimensional inputs, Perceiver uses an asymmetric cross attention method to distill input information into low-dimensional latent bottleneck features. Once the features are distilled in a compact and fixed-dimensional form, regular Transformer blocks are applied in the latent space. The original Perceiver model shows performance competitive to ResNets and ViTs on image classification and can process 3D data, audio, images, video or their combinations. However, this model can only generate fixed outputs *e.g.*, class probabilities. A recent improvement called Perceiver IO [275] aims to learn models with both flexible inputs as well as arbitrary sized outputs. This allows application to problems which demand structured outputs such as natural language tasks and visual comprehension. While these models avoid modality dependent architectural choices, the learning itself still involves modality dependent choices *e.g.*, specific augmentations or positional encodings. An interesting and open future direction is to achieve total modality-agnosticism in the learning pipeline.

5 CONCLUSION

Attention has played a key role in delivering efficient and accurate computer vision systems, while simultaneously providing insights into the function of deep neural networks. This survey reviews the self-attention approaches and specifically focuses on the Transformer and bidirectional encoding architectures that are built on the principle of self-attention. We first cover fundamental concepts pertaining to self-attention architectures and later provide an in-depth analysis of competing approaches for a broad range of computer vision applications. Specifically, we include state of the art self-attention models for image recognition, object detection, semantic and instance segmentation, video analysis and classification, visual question answering, visual commonsense reasoning, image captioning, vision-language navigation, clustering, few-shot learning, and 3D data analysis. We systematically highlight the key strengths and limitations of the existing methods and particularly elaborate on the important future research directions. With its specific focus on computer vision tasks, this survey provides a unique view of the recent progress in self-attention and Transformer-based methods. We hope this effort will drive further interest in the vision community to leverage the potential of Transformer models and improve on their current limitations *e.g.*, reducing their carbon footprint.

ACKNOWLEDGMENTS

The authors would like to thank Tim Prangemeier (TU Darmstadt), Luowei Zhou (Microsoft Research), Jason Corso (University of Michigan), Pichao Wang (Alibaba Group), Yuqing Wang (Meituan), Alex Meinke (Uni-Tuebingen), Irwan Bello (Google Brain) and Manoj Kumar (Google Brain) for their helpful feedback on the survey. We would also like to thank Mohamed Afham for his help with a figure.

REFERENCES

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.
- [2] M. Ott, S. Edunov, D. Grangier, and M. Auli, “Scaling neural machine translation,” in *WMT*, 2018.

- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [4] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," tech. rep., OpenAI, 2018.
- [5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," tech. rep., OpenAI, 2019.
- [6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.
- [7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.
- [9] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," *arXiv preprint arXiv:2006.16668*, 2020.
- [10] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*.
- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020.
- [13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.
- [14] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.
- [15] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *CVPR*, 2019.
- [16] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *CVPR*, 2020.
- [17] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *ICCV*, 2019.
- [18] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *CVPR*, 2019.
- [19] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," *arXiv preprint arXiv:2012.00364*, 2020.
- [20] A. Ramesh, M. Pavlov, G. Goh, and S. Gray, "DALL·E: Creating images from text," tech. rep., OpenAI, 2021.
- [21] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *EMNLP-IJCNLP*, 2019.
- [22] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "VL-BERT: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.
- [23] X. Wang, C. Yeshwanth, and M. Nießner, "SceneFormer: Indoor scene generation with transformers," *arXiv preprint arXiv:2012.09793*, 2020.
- [24] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," in *ICLR*, 2021.
- [25] C. Doersch, A. Gupta, and A. Zisserman, "CrossTransformers: spatially-aware few-shot transfer," *NeurIPS*, 2020.
- [26] H.-J. Ye, H. Hu, D.-C. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *CVPR*, 2020.
- [27] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," *arXiv preprint arXiv:1904.02874*, 2019.
- [28] A. de Santana Correia and E. L. Colombini, "Attention, please! a survey of neural attention models in deep learning," *arXiv preprint arXiv:2103.16775*, 2021.
- [29] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.
- [30] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*. MIT press, 2017.
- [31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.
- [32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.
- [33] D. Hu, "An introductory survey on attention mechanisms in nlp problems," in *IntelliSys*, 2019.
- [34] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *arXiv preprint arXiv:2009.06732*, 2020.
- [35] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021.
- [36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.
- [37] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *arXiv preprint arXiv:2104.13840*, 2021.
- [38] S. Wang, B. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.
- [39] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019.
- [40] J. Pérez, J. Marinković, and P. Barceló, "On the turing completeness of modern neural network architectures," in *International Conference on Learning Representations*, 2018.
- [41] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "On the relationship between self-attention and convolutional layers," in *International Conference on Learning Representations*, 2019.
- [42] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.
- [43] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "UNITER: Universal image-text representation learning," in *ECCV*, 2020.
- [44] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020.
- [45] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," *arXiv preprint arXiv:2012.09760*, 2020.
- [46] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.
- [47] "Revisiting the unreasonable effectiveness of data." <https://ai.googleblog.com/2017/07/revisiting-unreasonable-effectiveness.html>. Accessed: 2020-12-31.
- [48] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *TPAMI*, 2020.
- [49] X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *arXiv preprint arXiv:2006.08218*, 2020.
- [50] "Aaai 2020 keynotes turing award winners event." <https://www.youtube.com/watch?v=UX8OubxsY8w>. Accessed: 2020-12-31.
- [51] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *ECCV*, 2016.
- [52] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.
- [53] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.
- [54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.
- [55] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised representation learning for remote sensing image classification," *GRSL*, 2017.
- [56] U. Ahsan, R. Madhok, and I. Essa, "Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition," in *WACV*, 2019.
- [57] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *ECCV*, 2016.

- [58] D. Kim, D. Cho, D. Yoo, and I. S. Kweon, "Learning image representations by completing damaged jigsaw puzzles," *WACV*, 2018.
- [59] L. Jing, X. Yang, J. Liu, and Y. Tian, "Self-supervised spatiotemporal feature learning via video rotation prediction," *arXiv preprint arXiv:1811.11387*, 2018.
- [60] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang, "Unsupervised representation learning by sorting sequences," in *ICCV*, 2017.
- [61] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *ECCV*, 2016.
- [62] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in *CVPR*, 2018.
- [63] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," in *Arxiv preprint arXiv:1908.03557*, 2019.
- [64] B. Korbar, D. Tran, and L. T., "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018.
- [65] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017.
- [66] N. Sayed, B. Brattoli, and B. Ommer, "Cross and learn: Cross-modal self-supervision," in *GCPR*, 2018.
- [67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [68] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [69] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *CVPR*, 2005.
- [70] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.
- [71] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.
- [72] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *ICCV*, 2019.
- [73] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
- [74] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017.
- [75] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.
- [76] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *TPAMI*, 2018.
- [77] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, 2009.
- [78] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *ICCV*, 2019.
- [79] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in *ICCV*, 2019.
- [80] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *NAACL*, 2018.
- [81] N. Parmar, P. Ramachandran, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *NeurIPS*, 2019.
- [82] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *CVPR*, 2020.
- [83] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
- [84] M. M. Naseer, S. H. Khan, M. H. Khan, F. S. Khan, and F. Porikli, "Cross-domain transferability of adversarial perturbations," in *NeurIPS*, 2019.
- [85] M. Naseer, K. Ranasinghe, S. Khan, F. S. Khan, and F. Porikli, "On improving adversarial transferability of vision transformers," *arXiv preprint arXiv:2106.04169*, 2021.
- [86] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *CVPR*, 2020.
- [87] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.
- [88] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021.
- [89] Z. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, and J. Feng, "All tokens matter: Token labeling for training better vision transformers," *arXiv preprint arXiv:2104.10858*, 2021.
- [90] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.
- [91] A. El-Nouby, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek, and H. Jegou, "Xcit: Cross-covariance image transformers," 2021.
- [92] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," 2021.
- [93] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.
- [94] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," 2021.
- [95] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," 2021.
- [96] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," *arXiv preprint arXiv:2103.15808*, 2021.
- [97] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvtv2: Improved baselines with pyramid vision transformer," 2021.
- [98] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," 2021.
- [99] W. Wang, L. Yao, L. Chen, D. Cai, X. He, and W. Liu, "Crossformer: A versatile vision transformer based on cross-scale attention," *arXiv preprint arXiv:2108.00154*, 2021.
- [100] C.-F. Chen, R. Panda, and Q. Fan, "Regionvit: Regional-to-local attention for vision transformers," 2021.
- [101] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021.
- [102] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," *ICCV 2021*, 2021.
- [103] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
- [104] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *arXiv preprint arXiv:2103.14899*, 2021.
- [105] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," *arXiv preprint arXiv:2103.11816*, 2021.
- [106] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021.
- [107] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. V. Gool, "Localityvit: Bringing locality to vision transformers," 2021.
- [108] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jegou, and M. Douze, "Levit: A vision transformer in convnet's clothing for faster inference," 2021.
- [109] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [110] Q. Zhang and Y. Yang, "Rest: An efficient transformer for visual recognition," *arXiv preprint arXiv:2105.13677*, 2021.
- [111] Z. Zhang, H. Zhang, L. Zhao, T. Chen, and T. Pfister, "Aggregating nested transformers," in *arXiv preprint arXiv:2105.12723*, 2021.
- [112] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," 2021.
- [113] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, and C. Shen, "Conditional positional encodings for vision transformers," 2021.
- [114] Y. Liu, G. Sun, Y. Qiu, L. Zhang, A. Chhatkuli, and L. Van Gool, "Transformer in convolutional neural networks," *arXiv preprint arXiv:2106.03180*, 2021.

- [115] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised visual transformers," *arXiv e-prints*, pp. arXiv-2104, 2021.
- [116] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.
- [117] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.
- [118] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021.
- [119] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al., "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.
- [120] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.
- [121] C. Li, J. Yang, P. Zhang, M. Gao, B. Xiao, X. Dai, L. Yuan, and J. Gao, "Efficient self-supervised vision transformers for representation learning," *arXiv preprint arXiv:2106.09785*, 2021.
- [122] Y. Wang, X. Zhang, T. Yang, and J. Sun, "Anchor detr: Query design for transformer-based detector," 2021.
- [123] T. Chen, S. Saxena, L. Li, D. J. Fleet, and G. Hinton, "Pix2seq: A language modeling framework for object detection," 2021.
- [124] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," 2021.
- [125] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *TPAMI*, 2016.
- [126] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.
- [127] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *ICCV*, 2017.
- [128] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016.
- [129] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.
- [130] T. Prangemeier, C. Reich, and H. Koeppl, "Attention-based transformers for instance segmentation of cells in microstructures," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 700–707, IEEE, 2020.
- [131] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017.
- [132] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.
- [133] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," *arXiv preprint arXiv:2003.07853*, 2020.
- [134] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021.
- [135] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," 2021.
- [136] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *CVPR*, 2019.
- [137] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kortscheder, "The mapillary vistas dataset for semantic understanding of street scenes," in *ICCV*, 2017.
- [138] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.
- [139] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *ECCV*, 2016.
- [140] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016.
- [141] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *EMNLP*, 2014.
- [142] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *ICML*, 2018.
- [143] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *ICML*, 2020.
- [144] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," *arXiv:2012.09841*, 2020.
- [145] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two transformers can make one strong gan," 2021.
- [146] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, F. S. Khan, and M. Shah, "Handwriting transformers," *arXiv preprint arXiv:2104.03964*, 2021.
- [147] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., "Conditional image generation with pixelcnn decoders," in *NeurIPS*, 2016.
- [148] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., 2009.
- [149] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *AISTATS*, 2011.
- [150] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.
- [151] P. Bachman, R. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," in *NeurIPS*, 2019.
- [152] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Es-lami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.
- [153] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*, 2019.
- [154] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A guide to convolutional neural networks for computer vision," *Synthesis Lectures on Computer Vision*, 2018.
- [155] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [156] C. Gao, Y. Chen, S. Liu, Z. Tan, and S. Yan, "Adversarialnas: Adversarial neural architecture search for gans," in *CVPR*, pp. 5680–5689, 2020.
- [157] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *CVPR*, pp. 8110–8119, 2020.
- [158] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016.
- [159] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.
- [160] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *TPAMI*, 2018.
- [161] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *CVPR*, 2018.
- [162] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [163] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *NeurISP*, 2019.
- [164] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCVW*, 2021.
- [165] Z. Wang, X. Cun, J. Bao, and J. Liu, "Uformer: A general u-shaped transformer for image restoration," *arXiv preprint arXiv:2106.03106*, 2021.
- [166] Z. Lu, H. Liu, J. Li, and L. Zhang, "Efficient transformer for single image super-resolution," *arXiv preprint arXiv:2108.11084*, 2021.
- [167] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.
- [168] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *CVPR*, 2019.
- [169] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *ECCV*, 2020.
- [170] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017.

- [171] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *CVPR*, 2017.
- [172] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. Huang, "Image super-resolution via dual-state recurrent networks," in *CVPR*, 2018.
- [173] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *TPAMI*, 2020.
- [174] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "ESRGAN: enhanced super-resolution generative adversarial networks," in *ECCVW*, 2018.
- [175] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "SRFEAT: Single image super-resolution with feature discrimination," in *ECCV*, 2018.
- [176] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," in *ICCV*, 2017.
- [177] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017.
- [178] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.
- [179] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," *arXiv preprint arXiv:1912.12180*, 2019.
- [180] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, and M. Zhou, "Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training," in *AAAI*, 2020.
- [181] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.
- [182] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, and Y. Song, "Parameter efficient multimodal transformers for video representation learning," *arXiv preprint arXiv:2012.04124*, 2020.
- [183] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "VQA: Visual question answering," in *ICCV*, 2015.
- [184] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *CVPR*, 2019.
- [185] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *ECCV*, 2018.
- [186] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," *arXiv preprint arXiv:1811.00491*, 2018.
- [187] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv:1907.06987*, 2019.
- [188] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
- [189] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.
- [190] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *ECCV*, 2016.
- [191] H. Tan and M. Bansal, "Vokenization: Improving language understanding with contextualized, visual-grounded supervision," in *EMNLP*, 2020.
- [192] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *CVPR*, 2020.
- [193] A. Majumdar, A. Shrivastava, S. Lee, P. Anderson, D. Parikh, and D. Batra, "Improving vision-and-language navigation with image-text pairs from the web," *arXiv preprint arXiv:2004.14973*, 2020.
- [194] K. Chen, J. K. Chen, J. Chuang, M. Vázquez, and S. Savarese, "Topological planning with transformers for vision-and-language navigation," *arXiv preprint arXiv:2012.05292*, 2020.
- [195] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," *Image*, vol. 2, p. T2, 2021.
- [196] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018.
- [197] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *AAAI*, vol. 34, pp. 13041–13049, 2020.
- [198] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," *arXiv preprint arXiv:1906.05743*, 2019.
- [199] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," in *EMNLP*, 2019.
- [200] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.
- [201] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NeurIPS*, 2011.
- [202] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588, 2021.
- [203] A. Kamath, M. Singh, Y. LeCun, I. Misra, G. Synnaeve, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," *arXiv preprint arXiv:2104.12763*, 2021.
- [204] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "Transvg: End-to-end visual grounding with transformers," 2021.
- [205] M. Li and L. Sigal, "Referring transformer: A one-step approach to multi-task visual grounding," *arXiv preprint arXiv:2106.03089*, 2021.
- [206] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Visual grounding with transformers," *arXiv preprint arXiv:2105.04281*, 2021.
- [207] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "COOT: Cooperative hierarchical transformer for video-text representation learning," *arXiv preprint arXiv:2011.00597*, 2020.
- [208] H. Seong, J. Hyun, and E. Kim, "Video multitask transformer network," in *ICCV Workshops*, pp. 0–0, 2019.
- [209] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," *arXiv preprint arXiv:2011.14503*, 2020.
- [210] L. Zhou, Y. Zhou, J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *CVPR*, 2018.
- [211] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," *arXiv preprint arXiv:2102.00719*, 2021.
- [212] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," *arXiv preprint arXiv:2103.15691*, 2021.
- [213] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.
- [214] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles, "Dense-captioning events in videos," in *ICCV*, pp. 706–715, 2017.
- [215] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI*, vol. 32, 2018.
- [216] C. Plizzari, M. Cannici, and M. Matteucci, "Spatial temporal transformer network for skeleton-based action recognition," *arXiv preprint arXiv:2008.07404*, 2020.
- [217] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016.
- [218] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding," *TPAMI*, 2019.
- [219] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," 2021.
- [220] J. Wang, G. Bertasius, D. Tran, and L. Torresani, "Long-short temporal contrastive learning of video transformers," *arXiv preprint arXiv:2106.09212*, 2021.
- [221] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *ICCV*, pp. 5188–5197, 2019.
- [222] G. Bertasius and L. Torresani, "Classifying, segmenting, and tracking object instances in video with mask propagation," in *CVPR*, pp. 9739–9748, 2020.
- [223] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, et al., "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *ICLR*, 2020.

- [224] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
- [225] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *NeurIPS*, 2017.
- [226] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," 2020.
- [227] H. Edwards and A. Storkey, "Towards a neural statistician," *arXiv preprint arXiv:1606.02185*, 2016.
- [228] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *ICML*, 2019.
- [229] J. Lee, Y. Lee, and Y. W. Teh, "Deep amortized clustering," *arXiv preprint arXiv:1909.13433*, 2019.
- [230] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," *arXiv preprint arXiv:2012.09164*, 2020.
- [231] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *arXiv preprint arXiv:2012.09688*, 2020.
- [232] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *CVPR*, 2015.
- [233] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.
- [234] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *TPAMI*, 2013.
- [235] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018.
- [236] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. Brox, "FreiHAND: A dataset for markerless capture of hand pose and shape from single rgb images," in *ICCV*, 2019.
- [237] "OpenAI's GPT-3 language model: A technical overview." <https://lambdalabs.com/blog/demystifying-gpt-3/>. Accessed: 2020-12-31.
- [238] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," 2021.
- [239] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, 2014.
- [240] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017.
- [241] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *ICCV*, 2015.
- [242] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *NeurIPS*, 2017.
- [243] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," *arXiv preprint arXiv:2103.17239*, 2021.
- [244] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.
- [245] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," *arXiv:1904.10509*, 2019.
- [246] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *ICLR*, 2020.
- [247] I. Bello, "Lambdanetworks: Modeling long-range interactions without attention," in *International Conference on Learning Representations*, 2021.
- [248] A. Vyas, A. Katharopoulos, and F. Fleuret, "Fast transformers with clustered attention," *NeurIPS*, 2020.
- [249] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2t: Pyramid pooling transformer for scene understanding," *arXiv preprint arXiv:2106.12011*, 2021.
- [250] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12894–12904, 2021.
- [251] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," *arXiv preprint arXiv:2107.00652*, 2021.
- [252] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh, "Nystr\\" omformer: A nystr\\" om-based algorithm for approximating self-attention," in *AAAI*, 2021.
- [253] Y. Tay, D. Bahri, D. Metzler, D. Juan, Z. Zhao, and C. Zheng, "Synthesizer: Rethinking self-attention in transformer models," in *ICML*, 2021.
- [254] H. Peng, N. Pappas, D. Yogatama, R. Schwartz, N. A. Smith, and L. Kong, "Random feature attention," in *ICLR*, 2021.
- [255] K. Choromanski, V. Likhoshesterov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al., "Rethinking attention with performers," in *ICLR*, 2021.
- [256] Y. Tay, D. Bahri, L. Yang, D. Metzler, and D.-C. Juan, "Sparse sinkhorn attention," in *ICML*, 2020.
- [257] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform resnets without pretraining or strong data augmentations," *arXiv preprint arXiv:2106.01548*, 2021.
- [258] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.
- [259] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- [260] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," *arXiv:2102.04378*, 2021.
- [261] D. R. So, C. Liang, and Q. V. Le, "The evolved transformer," 2019.
- [262] H. Wang, Z. Wu, Z. Liu, H. Cai, L. Zhu, C. Gan, and S. Han, "Hat: Hardware-aware transformers for efficient natural language processing," 2020.
- [263] M. Chen, H. Peng, J. Fu, and H. Ling, "Autoformer: Searching transformers for visual recognition," *arXiv preprint arXiv:2107.00651*, 2021.
- [264] C. Li, T. Tang, G. Wang, J. Peng, B. Wang, X. Liang, and X. Chang, "Bossnas: Exploring hybrid cnn-transformers with block-wisely self-supervised neural architecture search," *arXiv preprint arXiv:2103.12424*, 2021.
- [265] B. Chen, P. Li, C. Li, B. Li, L. Bai, C. Lin, M. Sun, W. Ouyang, et al., "Glit: Neural architecture search for global and local image transformer," *arXiv preprint arXiv:2107.02960*, 2021.
- [266] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," *arXiv preprint arXiv:2105.10497*, 2021.
- [267] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," *arXiv preprint arXiv:1905.09418*, 2019.
- [268] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.
- [269] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," *arXiv preprint arXiv:2012.09838*, 2020.
- [270] B. Li, S. Pandey, H. Fang, Y. Lyv, J. Li, J. Chen, M. Xie, L. Wan, H. Liu, and C. Ding, "FTRANS: energy-efficient acceleration of transformers using fpga," in *ISLPED*, 2020.
- [271] G. Bender, P.-J. Kindermans, B. Zoph, V. Vasudevan, and Q. Le, "Understanding and simplifying one-shot architecture search," in *ICML*, 2018.
- [272] Z. Guo, X. Zhang, H. Mu, W. Heng, Z. Liu, Y. Wei, and J. Sun, "Single path one-shot neural architecture search with uniform sampling," *arXiv preprint arXiv:1904.00420*, 2019.
- [273] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, "Efficient neural architecture search via parameter sharing," in *ICML*, 2018.
- [274] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver: General perception with iterative attention," *arXiv preprint arXiv:2103.03206*, 2021.
- [275] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, et al., "Perceiver io: A general architecture for structured inputs & outputs," *arXiv preprint arXiv:2107.14795*, 2021.