Attention mechanisms and deep learning for machine vision: A survey of the state of the art

Abdul Mueed Hafiz (mueedhafiz@uok.edu.in)
University of Kashmir https://orcid.org/0000-0002-2266-3708
Shabir Ahmad Parah
University of Kashmir
Rouf Ul Alam Bhat
University of Kashmir

Research Article

Keywords: Attention, vision transformers, CNNs, deep learning, machine vision

Posted Date: June 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-510910/v1

Noname manuscript No.
(will be inserted by the editor)

Attention mechanisms and deep learning for machine vision: A survey of the state of the art

Abdul Mueed Hafiz · Shabir Ahmad Parah · Rouf Ul Alam Bhat

Received: date / Accepted: date

Abstract With the advent of state of the art nature-inspired pure attention based models i.e. transformers, and their success in natural language processing (NLP), their extension to machine vision (MV) tasks was inevitable and much felt. Subsequently, vision transformers (ViTs) were introduced which are giving quite a challenge to the established deep learning based machine vision techniques. However, pure attention based models/architectures like transformers require huge data, large training times and large computational resources. Some recent works suggest that combinations of these two varied fields can prove to build systems which have the advantages of both these fields. Accordingly, this state of the art survey paper is introduced which hopefully will help readers get useful information about this interesting and potential research area. A gentle introduction to attention mechanisms is given, followed by a discussion of the popular attention based deep architectures. Subsequently, the major categories of the intersection of attention mechanisms and deep learning for machine vision (MV) based are discussed. Afterwards, the major algorithms, issues and trends within the scope of the paper are discussed.

Keywords Attention · vision transformers · CNNs · deep learning · machine vision

Abdul Mueed Hafiz
Dept of Electronics & Communication Engineering, Institute of Technology, University of Kashmir (Zakura Campus), Srinagar, J&K, 190006 India
Tel.: +91-7006474254
E-mail: mueedhafiz@uok.edu.in

Shabir Ahmad Parah
Department of Electronics and Instrumentation Technology, University of Kashmir (Main Campus), Srinagar, J&K, 190006 India

Rouf Ul Alam Bhat
Dept of Electronics & Communication Engineering, Institute of Technology, University of Kashmir (Zakura Campus), Srinagar, J&K, 190006 India

1 Introduction

Recently attention-based mechanisms like transformers [93] have been successfully applied to various machine vision tasks by using them as vision transformers (ViTs) [20] in image recognition [90], object detection [8,119], segmentation[112], image super-resolution [109], video understanding [86,26], image generation [10], text-image synthesis [75] and visual question answering [87,85], among others [97,50, 18,111] achieving at par as well as even better results as compared to the established CNN models [45]. However, transformers have various issues like being 'data-hungry' and requiring large training times. Deep learning [54,27,82] based convolutional neural networks (CNNs) [55,56] on the other hand do not have such problems significantly. Accordingly, techniques have emerged which are at the intersection of pure attention based models and the established pure CNNs which have best of the both features. Machine vision (MV) has also benefitted from this merger of the two important vision models viz. ViTs and CNNs. In the this section we will discuss the source of power of ViTs and transformers in general i.e. attention and its types [45] briefly for the readers to have an idea of the new type of machine vision (MV) models i.e. ViTs.

1.1 Self-attention

For a given a sequence of elements, the self-attention process gives a measurable estimate of the relevance of one element others. For example, which elements like words can come together in a sequence like a sentence. The self-attention process is an important unit of attention-based models like transformers, that models the dependencies among all elements of the sequence for formal/structured prediction applications. Plainly stated, a self-attention model layer assigns a value to every element in a structure/sequence by combining information globally from the input vector/sequence.

Denoting a sequence of n entities ($\mathbf{x}_1$, $\mathbf{x}_2$, ... $\mathbf{x}_n$) by $\mathbf{X} \in \mathbb{R}^{n \times d}$, d being the dimension which embeds dependency of every element. The purpose of self-attention is capturing the dependency between all n elements after encoding every element inside the overall contextual knowledge. This process is achieved by the definition of 3 weight matrices which have to be learnt for transforming: Queries $\left(\mathbf{W}^Q \in \mathbb{R}^{n \times d_q}\right)$, Keys $\left(\mathbf{W}^K \in \mathbb{R}^{n \times d_k}\right)$ and Values $\left(\mathbf{W}^V \in \mathbb{R}^{n \times d_v}\right)$. First the input vector $\mathbf{X}$ is projected to the 3 weight matrices for obtaining $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$ and $\mathbf{V} = \mathbf{X}\mathbf{W}^V$. The output $\mathbf{Z} \in \mathbb{R}^{n \times d_v}$ in the self-attention layer is next expressed as,

$$Z = \textbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_q}}\right)\mathbf{V}. \tag{1}$$

For a certain element in the vector/sequence, the self-attention mechanism fundamentally finds the dot product of query with all the keys, this product being subsequently normalized by the softmax function for obtaining the attention-map scores. Every element now assumes the value of the weighted summation for all elements inside the vector/sequence, wherein all weights are equal to the attention map scores.

1.2 Masked self-attention

The self-attention layer applies to every element/entity. For the transformer [93] having been trained for prediction of the next entity in the vector/sequence, the self-attention units inside the decoder are then masked for prevention of their application to the entities coming in future. This technique is achieved by calculating the element-wise product with a mask $\mathbf{M} \in \mathbb{R}^{n \times n}$, where $\mathbf{M}$ is the upper triangular matrix. Thus masked self-attention is calculated as,

$$\textbf{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_q}} \circ \mathbf{M}\right), \tag{2}$$

where $\circ$ is the Hadamard product. During prediction of an element in the vector/sequence, the attention map scores of the future elements are set to 0 in the masked self-attention.

1.3 Multi-head attention

For encapsulation of various complicated dependencies between various elements / entities in the vector/sequence, the multi-head attention process consists of multiple self-attention units with $h = 8$ inside the original transformer architecture [93]. Every unit contains its own learnable weight-matrices $\left\{\mathbf{W}^{Q_i}, \mathbf{W}^{K_i}, \mathbf{W}^{V_i}\right\}$, where *i = 0, 1, 2, ... (h - 1)*. For a particular input $\mathbf{X}$, outputs of h self-attention units in the multi-head attention process are combined into one matrix $[\mathbf{Z}_0, \mathbf{Z}_1, \ldots \mathbf{Z}_{\mathrm{h}-1}] \in \mathbb{R}^{n \times h \times d_v}$ and are subsequently projected to another weight matrix $\mathbf{W} \in \mathbb{R}^{h \cdot d_v \times d}$.

The notable difference of the self-attention process with the convolutional operation is that every weight is dynamically computed as against static weights which remain fixed for various inputs as for convolution. Also that the self-attention process is invariable to permutation and change for different number of inputs with the result that it has a convenient operation over irregularity as against the convolutional operator which needs a grid array. See 1 for illustration of these concepts.

2 Attention based deep learning architectures

In this section, some common deep learning architectures of deep attention models are discussed [96] and a graphical illustration is presented in 2. The architectures of the prevalent deep attention based models are categorized into the following important classes as given below:

1. Single channel model
2. Multi-channel model feeding on multi-scale data
3. Skip-layer model
4. Bottom-up/ top-down model
5. Skip-layer model with multi-scale saliency single network

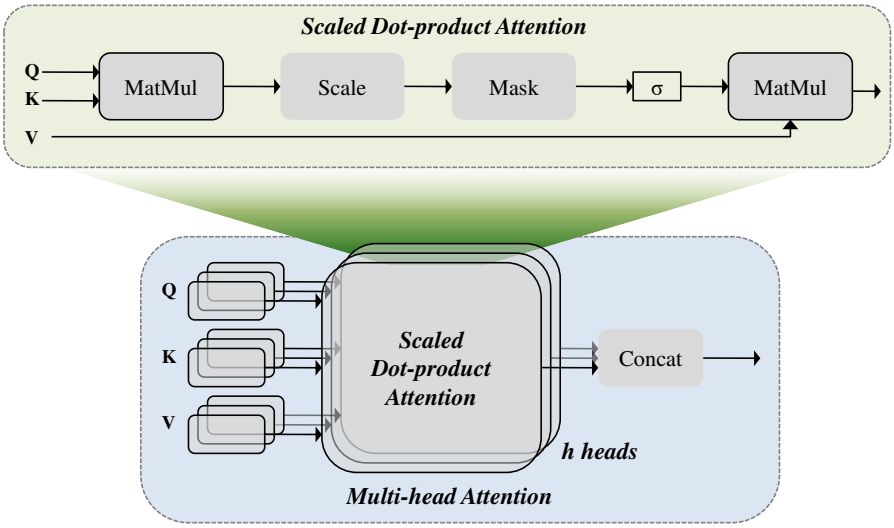

Fig. 1: Illustration of various attention mechanisms [45]

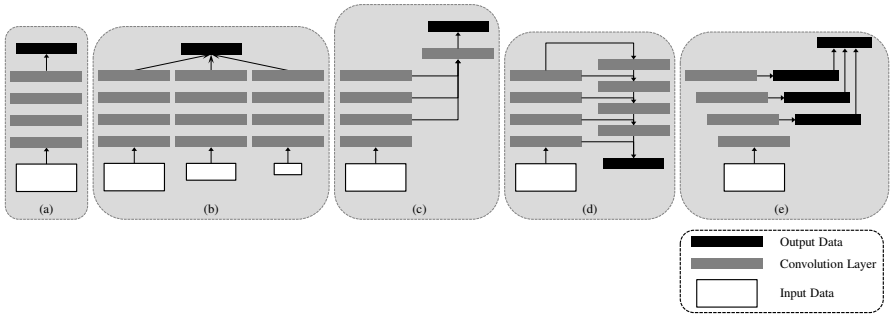

Fig. 2: (a)-(c) Depiction of 3 common classes of deep learning configurations used in attention prediction: (a) single-channel model configuration, (b) multi-channel model configuration with multi-scale input data, and (c) skip-layer model configuration. (d) bottom-up/top-down model configuration used in attention-based object segmentation and instance segmentation. (e) modified skip-layer model using multi-scale attention information in a single network. [96]

2.1 Single-channel model

As demonstrated in Figure 2(a), the single channel model is the predominant configuration of various CNN-based attention models also being used by many attention-based works [48,43,49,72]. Almost all the other types of CNN configurations can be considered as variants of the single channel model. It has been demonstrated that attention cues on various levels and scales are vital for attention [108]. Using multi-

scale features of CNNs into attention-based models is an obvious choice. In the next type of single channel model, namely multi-channel model, the changes are done along this line.

2.2 Multi-channel model

Some implementations of this model include [41,116,61,67]. The basic concept in the multi-channel model is shown in Figure 2(b). This type of model learns multi-scale attention information by training multiple models with multi-scale data inputs. The multiple model channels are in parallel and can have varying configurations with different scales. As shown in [105], input data is fed via multiple channels simultaneously, and then the features from different channels are fused and fed into a unified output layer for producing the final attention map. We observe that in the multi-channel model, multi-scale learning takes place outside the individual models. In the next configuration discussed, the multi-scale learning is inside the model, and this is achieved by combining feature maps from various convolutional layer hierarchies.

2.3 Skip-layer model

A common skip-layer model is shown in Figure 2(c) being used in [51,52,14]. Instead of learning from many parallel channels on multiple-scale images, the skip-layer model learns multi-scale feature maps inside a primary channel. Multi-scale outputs are learned from various layers with increasingly larger reception fields and down-sampling ratios. Next, these outputs are fused for outputting final attention map.

2.4 Bottom-up/top-down model

This relatively newer model configuration called top-down/bottom-up model has been used in attention-based object segmentation [110] and also in instance segmentation [73,31,29]. The architecture of the model is shown in Figure 2(d), wherein segmentation feature maps are first obtained by common bottom-up convolution techniques, and next a top-down refinement is done for fusing the data from deep to shallow layers into the mask. The main motivation behind this configuration is to produce high-fidelity segmentation masks because deep CNN layers lose fine image detail. The bottom-up/top-down model is like a type of skip-layer model since different layers are connected to each other.

2.5 Skip-layer Model with Multi-scale Saliency Single Network

This model [96] shown in Fig. 2(e), is inspired by the model in [105] and the deeply-supervised model in [57]. The model uses multi-scale and multi-level attention-based information from various layers, and learns via the deeply supervised technique. An important difference between this model and the previous models is that the former

provides combined straightforward supervision of the hidden layers instead of the common approach of supervising only the last output layer and then propagating the supervised output back to the previous layers. It uses the merit of the skip-layer model (Figure 2(c)) which does not learn from multiple model channels with multi-scale input data. Also, it is lighter than the multi-channel model (Figure 2(b)) and bottom-up/top-down model (Figure 2(d)). It has been found that the bottom-up/top-down model faces training difficulties while as the deeply supervised model shows high training efficiency.

In the next section we turn to the categorization of various techniques of attention mechanisms and deep learning in machine vision, and discuss each category in detail.

3 Attention and deep learning in machine vision: Broad categories

In this section, we discuss category-wise the various techniques of attention mechanisms and deep learning applied to machine vision. Three broad categories are:

1. Attention-based CNNs
2. CNN transformer pipelines
3. Hybrid transformers

These categories are discussed in the following sub-sections one by one. First we discuss attention-based CNNs in the following subsection.

3.1 Attention-based CNNs

Recently attention mechanisms have been applied in deep learning for machine vision applications, e.g. object detection [5,83,76], image captioning [106,113,3] and action recognition [81]. The central idea of the attention mechanisms is locating the most salient components of the feature maps in convolutional neural networks (CNNs) in a manner that the redundancy is removed for machine vision applications. Generally, attention is embedded in the CNN by using attention maps. Particularly the attention-based maps in [83,76,106,81] yield in a self learned manner having other information with weak supervision of the attention maps. Other techniques cited in literature [113,107] proceed by utilization of human attention data or guidance of the CNNs by focusing on the regions of interest (ROIs). In the following subsections, we proceed with discussing some noteworthy techniques in the general area of machine vision which use attention-based CNNs e.g. those used in image classification/retrieval, object detection, sign language recognition, denoising and facial expression recognition.

3.1.1 Image classification/retrieval and object detection

It is well established that attention contributes to human perception in an important manner [47,78,13]. One important characteristic of a human vision system is that it does not attempt to address the whole visual scene at one go. Instead, in the same,

a sequence of partial glimpses is exploited and focusing is done selectively on various parts for capturing the visual structure in a better manner [53]. Recently, several attempts have been made [95,38] for incorporation of attention processing mechanisms in order to improve the classification accuracy of CNNs on large scale classification tasks. Wang et al. [95] have proposed a residual attention network having an encoder-decoder style attention mechanism unit. By refinement of the features, the network gives good accuracy as well as shows robustness to noise. Without directly computing the three dimensional attention map, the process is decomposed such that it learns channel-attention and spatial-attention exclusively. The exclusive attention map generation technique for 3D features is computationally inexpensive and parameter restricted, and hence can be used as a plug and play unit for existing CNN networks. In their work [38], the authors have introduced a compact unit for exploitation of the relationship between various channels. In this 'Squeeze and Excitation' unit, the authors have used global average-pooling of feature maps for computation of each channel's attention. However, the authors of [101] show that the features used in [38] are suboptimal for inferring fine-channel attention. Accordingly the authors of [101] use max-pooled feature maps also. According to [101] in [38] spatial attention is missed which contributes in an important manner to deciding the focusing region as brought out in [11]. The authors of [101] thus proposed the convolutional block attention module (CBAM) for exploitation of both the spatial as well as channel-wise attention with the help of a robust network and proceed to verify that exploitation of both these mechanisms is better than use of only the channel-wise attention mechanism [38] by using it for image classification in ImageNet-1K dataset [15]. The authors of [101] experimentally demonstrate that their module is effective also in object detection tasks using two popular datasets viz. MS-COCO [66] and VOC [23]. They achieve impressive results by inserting their module in the pre-existing one-shot object detector [100] in the VOC-2007 testing set. 3 shows the CBAM for both channel and spatial-attention processes. Here we attempt to briefly explain the attention mechanism in CBAM.

For a given input feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ **,** CBAM [101] produces a one-dimensional attention map $\mathbf{M}_c \in \mathbb{R}^{C\times 1\times 1}$ and a two-dimensional spatial attention map $\mathbf{M}_s \in \mathbb{R}^{1\times H\times W}$ as shown in Figure 3. This attention mechanism operation can be put as:

$$\mathbf{F}^{'} = \mathbf{M}_c\left(\mathbf{F}\right) \bigotimes \mathbf{F}, \tag{3}$$

$$\mathbf{F}^{''} = \mathbf{M}_s\left(\mathbf{F}'\right) \bigotimes \mathbf{F}'. \tag{4}$$

where $\bigotimes$ is the multiplication operator for elements.

Channel attention is mathematically computed as follows:

$$\begin{aligned}\mathbf{M}_c\left(\mathbf{F}\right) &= \sigma\left(MLP\left(AvgPool\left(\mathbf{F}\right)\right) + MLP\left(MaxPool\left(\mathbf{F}\right)\right)\right) \\ &= \sigma\left(\mathbf{W}_1\left(\mathbf{W}_0\left(\mathbf{F}^c_{\text{avg}}\right)\right) + \mathbf{W}_1\left(\mathbf{W}_0\left(\mathbf{F}^c_{\max}\right)\right)\right)\end{aligned} \tag{5}$$

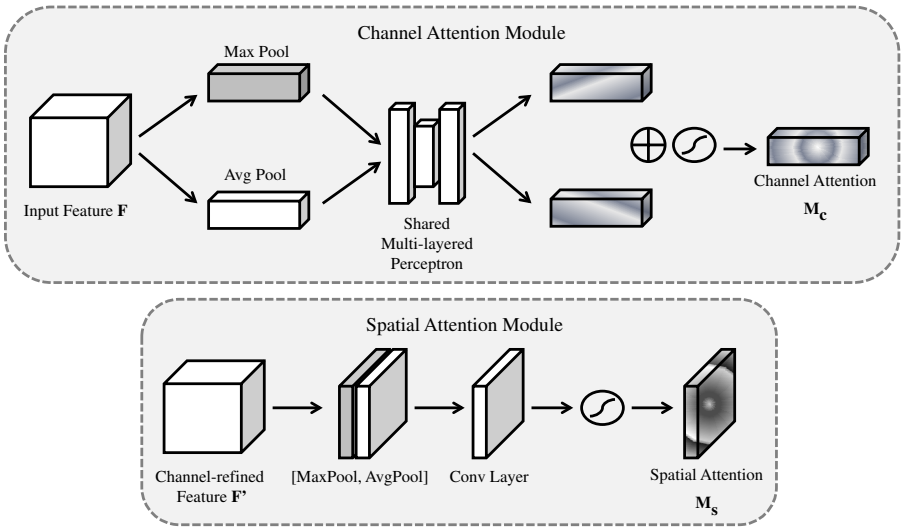

Fig. 3: Illustration of both attention sub-modules in CBAM [101]. As shown, the channel-wise sub-module utilizes the max-pooling of the feature output as well as the average-pooling of the feature output with the help of a shared network. On the other hand, the spatial-wise sub-module uses two identical feature outputs by pooling them along their channel axes and then forwarding them to the convolutional layer. [101]

where σ is the sigmoid function, $\mathbf{W}_0 \in \mathbb{R}^{C/r \times C}$ and $\mathbf{W}_1 \in \mathbb{R}^{C \times C/r}$. The Multi-layer Perceptron (MLP) weights, $\mathbf{W}_0$ and $\mathbf{W}_1$, are shared for both the inputs. $\mathbf{W}_0$ comes after the ReLU activation function.

$$\mathbf{M}_s(\mathbf{F}) = \sigma\left(f^{7\times7}\left([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})]\right)\right) = \sigma\left(f^{7\times7}\left(\left[\mathbf{F}^s_{\text{avg}}; \mathbf{F}^s_{\text{max}}\right]\right)\right) \quad (6)$$

where σ is the sigmoid function and $f^{7\times7}$ is the convolutional operator with a 7×7 filter.

The authors of [101] have used their technique for both image classification/retrieval on the ImageNet-1K dataset and object detection on both MS-COCO and VOC2007 datasets. The results obtained using their CBAM integrated networks outperform other contemporary networks. They also have demonstrated the superiority of their technique as compared to others also via grad-CAM [79] visualizations obtained on images from ImageNet validation set. 4 shows the same.

Another novel and related work in the area of image classification/retrieval by using attention-based CNNs is given by the authors of [63] for glaucoma detection from the area of medical image analysis [30]. They call their network attention-based CNN for glaucoma detection AG-CNN. It includes a novel attention-prediction sub-net along with other subnets. They achieve 'end to end' training on an attention-

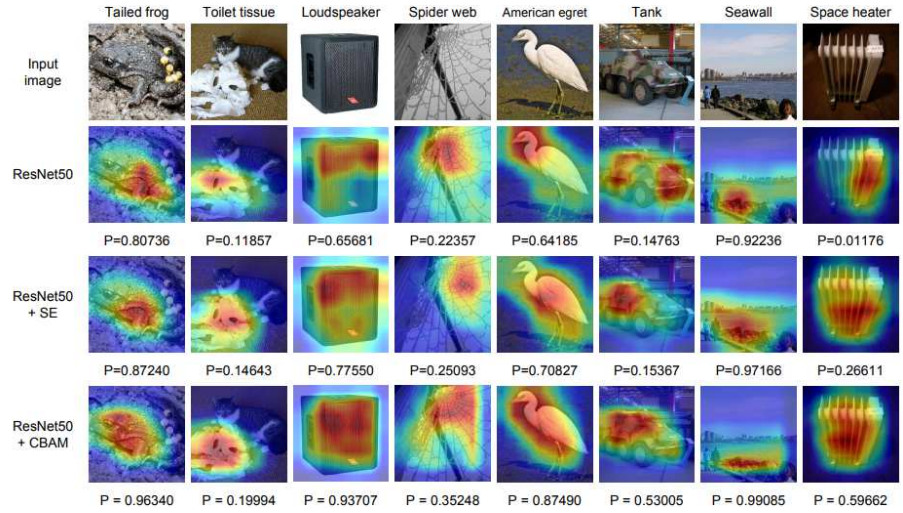

Fig. 4: Heat map visualizations using Grad-CAM [79]. The visualizations are shown for those of the CBAM-fitted CNNs, viz. {**ResNet50 + CBAM**}, baseline {**ResNet50**} [33], and Squeeze and Excitation method [38] (SE)-integrated architecture {**ResNet50 + SE**}. Grad-CAM visualization has been obtained with feature maps of last conv layer outputs. The GT label has been shown on top of every image, where P is the softmax score of every network for the GT category. (Reproduced by permission from publisher of [101])

based CNN architecture by supervision of the training through 3 separate loss functions based on: i) attention-prediction, ii) feature-visualization and iii) glaucoma-classification. Based on the work of authors in [42], the authors use the Kullback Leibler (KL) divergence function as an equivalent of the nature-inspired attention-loss $Loss_a$ given by:

$$\text{Loss}_a = \frac{1}{I \cdot J} \sum_{i=1}^{I} \sum_{j=1}^{J} A_{ij} \log \left(\frac{A_{ij}}{\hat{A}_{ij}} \right) \tag{7}$$

where $\hat{A}$ (with its elements $\hat{A}_{ij} \in [0,1]$) is the attention map, and, I and J are the attention-map length and width respectively. By incorporating these novel features, the authors of [63] demonstrate that their proposed AG-CNN technique significantly improves the state of the art in glaucoma detection.

For more interesting techniques on image classification/retrieval using attention-based CNNs the readers may refer to some of the recent outstanding works in this area as given in [24,32], etc.

3.1.2 Sign Language Recognition

Sign language recognition (SLR) is a valuable and challenging research area in machine vision related multimedia field. Conventionally, SLR relies on hand-crafted

features with low performance. In their novel work [40], the authors propose to use attention based 3D CNNs for SLR. Their model has 2 advantages. First, it learns spatial and temporal features from video frames without any pre-processing or prior knowledge. Attention mechanisms help the model to select the clues. During training for capturing the features, spatial attention is used in the model for focusing on the ROIs. After this, temporal attention is used for selection of the important motions for determining the action-class. Their method has been benchmarked on a self-made large Chinese SL dataset having 500 classes, and also on the ChaLearn14 benchmark [22]. The authors demonstrate that their technique outperforms other state of the art techniques on the datasets used. We discuss this interesting technique in more detail below.

The spatial attention map is calculated as follows. They use an attention-based mask for denotation of the value of each image pixel. Let $x_{i,k} \in \mathbb{R}^2$ denote the position of a viewpoint k in an image i, the value of the location $p \in \mathbb{R}^2$ inside the attention map $M_{i,k} \in \mathbb{R}^{w \times h}$ for k is given by:

$$M_{i,k}(p) = \exp\left(-\frac{\|p - x_{i,k}\|_2^2}{\sigma}\right), \tag{8}$$

where σ is experimentally chosen, and w and h are image dimensions. The attention mask is formed by aggregating the peaks of various viewpoints obtained previously with the help of a max operator,

$$M_i(p) = \max_k M_{i,k}(p). \tag{9}$$

Consequently the i^{th} attention weighed image I_i is the element-wise product given by,

$$I_i(p) = I_i(p) \times M_i(p). \tag{10}$$

Based on the video feature obtained above, the use a Support Vector Machine (SVM) based classifier [92] for classification by clubbing it to another temporal attention-based pipeline. As done earlier in [21], the features are fed to a bi-directional LSTM for generation of an attention vector $s \in \mathbb{R}^{8192}$. The features are also fed to a one-layer MLP which gives the hidden vector $H = \{h_1, h_2, \ldots, h_n\}$, $h_i \in \mathbb{R}^{8192}$. This vector is an integration of the sequence of clip features by attention pooling. This technique measures the value of each clip feature by determining its relation with the attention vector s. Finally, they combine the video and trajectory features and use softmax based classification. Although an effective technique, the authors still admit that the work focuses on isolated SLR. For dealing with continuous SLR, which translates a clip into a sentence, RNN based methods are going to give results as admitted by the authors of the above work, and they want to work in that direction.

3.1.3 Image denoising

Image denoising is a low-level machine vision (MV) task. Deep CNNs are quite popular in low-level MV. Research has been done to improve the performance in the area by using very deep networks. However, as the network depth increases, the effects of the shallow layers on deep layers decrease. Accordingly the authors of [89] have proposed an attention-based denoising CNN named ADNet featuring an attention block (AB). The AB has been used for fine extraction of the noise data hidden in complex backgrounds. This technique has been proved by the authors of [89] to be very effective for denoising images with complex noise e.g. real noise-induced images. Various experiments demonstrate that ADNet delivers very good performance for 3 tasks viz. denoising of synthetic images, denoising of real noisy images, and also blind denoising. Here, the AB guides the previous network section by using the current network section in order to learn the noise nature. This is particularly useful for unknown images having noise, i.e. real noisy images and blind denoising. The AB uses 2 successive steps for implementation of its attention mechanism. First a 1x1 convolution is done on the output from the 17^{th} CNN layer output in order to compress the feature map into a weight vector for adjustment of the previous section. Next, the weights thus obtained are used to multiply the feature map output of the 16^{th} CNN layer for extraction of more refined noise feature maps. It should be noted that inspired by this novel effort more complex attention mechanisms can be used along with more dedicated 'denoising' deep CNNs. The code of ADNet is available at: https://github.com/hellloxiaotian/ADNet.

3.1.4 Facial expression recognition

One hot topic in Machine Vision (MV) is facial expression recognition (FER) which can be used in various MV fields like human computer interaction (HCI), affective computing, etc. In their work [62], the authors have proposed an end to end CNN network featuring an attention mechanism for auto FER. It has 4 main parts viz. feature extraction unit, attention unit, reconstruction unit and classification unit. The attention mechanism incorporated guides the CNN for paying more attention to important features extracted from earlier unit. The authors have combined their LBP features and their attention mechanism for enhancing the attention mechanism for obtaining better performance. They have applied their technique to their own dataset and 4 others, i.e., JAFFE [70], CK+ [69], FER2013 [1] and Oulu-CASIA [115], and have experimentally demonstrated that their technique performs better than other contemporary techniques. The attention mechanism used in the work has been proved to be valuable in pixelwise MV tasks. Their attention unit consists of two branches. The first is used to obtain feature map F_p, and the second combines the LBP feature maps for obtaining the attention maps F_m. In the next step, the element wise multiplication is done for the attention maps F_m and the feature maps F_p to obtain the final feature maps F_m as:

$$F_{final} = F_p F_m \tag{11}$$

Supposing that input of previous layer in the second branch is f_m, then the attention maps F_m are given by:

$$F_m = sigmoid\,(W f_m + b) \tag{12}$$

where w and b are denotations for weights and bias of conv layer, respectively. The technique is suitable for 2D images and its architecture needs to be modified to extend its application to video, 3D facial data, depth-image data. The authors also state that they are considering using more robust and efficient machine learning (ML) techniques for enhancement of the architecture.

In another valuable work in the area of FER given in [65],the authors state that in spite of the fact that conventional FER systems are almost perfect for analyzing constrained poses however they cannot perform well for partially occluded poses which are common in the real world. Accordingly, they have proposed an attention-based CNN (ACNN) for perception of facial occlusion part which focuses on the highly discriminative unoccluded parts. Learning in their model is end to end. For various Regions of Interest (ROIs), they have introduced two types of ACNN viz. patch based type and global-local based type. The first type uses attention only for local patches in face regions. The second type combines local features at the patch level with global features at the imagelevel. Evaluation is done on their own face expression dataset having in-the-wild occlusions, 2 of the largest in-the-wild face expression datasets i.e. RAF-DB [64] and AffectNet [71] and many other datasets. They show experimentally that using ACNNs improves the FER performance wherein the ACNNs shift attention from occluded facial regions to others which are not. They also show that their ACNN outperforms other state of the art techniques on several important FER datasets. However, the technique relies on landmarks. The authors intend to address this issue, as according them, ACNNs rely on face landmark localization units. Hence ACNNs have to be made more robust for generation of attention maps without landmarks, and this is an open area for research.

In the next sub-section, we turn to another important category of techniques of attention mechanisms and deep learning in machine vision, namely CNN transformer pipelines.

3.2 CNN transformer Pipelines

In this sub-section, we discuss another important category of techniques of attention and deep learning in machine vision, namely the CNN transformer pipeline. Here a CNN is used to feed feature maps to a transformer, and acts like a teacher to the transformer, as will be discussed. The notable works falling under this category have been discussed below for each area of machine vision (MV).

3.2.1 Image recognition

Transformers are 'data-hungry' in nature. For example a large-scale dataset like ImageNet [15] is not sufficient to train a vision transformer from scratch. To address

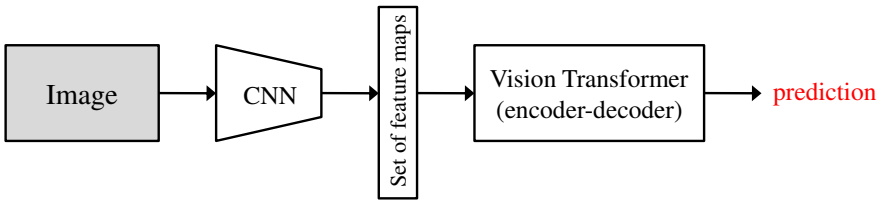

Fig. 5: Overview of the DETR pipeline. [8]

this issue, the work in [90] proposes to distill information from a teacher CNN to a student transformer, in turn allowing training of the transformer only on ImageNet sans additional data. The data-efficient image transformer (DeiT) [90] is a first in large scale image classification/retrieval without using a large-scale dataset like JFT [2]. DeiT shows that transformers (requiring very large amounts of training data) can also be trained successfully on medium-sized datasets (e.g., 1.2M images as against 100M+ images used in ViT [20]) with shorter training time. An important contribution of DeiT is its novel native distillation technique [36] which uses a teacher CNN (RegNetY-16GF [74]) whose outputs are fed to the transformer for training. The feature map outputs from the teacher CNN help the transformer (DeiT) in effectively finding important representations in input data images. The representations learned by DeiT are as good as top-performing CNNs like EfficientNet [88] and also are efficiently applicable to various downstream image recognition tasks.

3.2.2 Object detection

Like image classification/retrieval, transformers can be applied to image feature-map sets obtained from CNNs for precise object detection which involves prediction of object bounding boxes (BBoxes) and their corresponding category labels. In DETR [8], given spatial features obtained from a CNN backbone, the transformer encoder flattens the spatial axes along a single axis as shown in 5 which is feature map flattening from 3D to 1D. A sequence of features $(d \times n)$ is obtained with $d =$ feature dimension, and $n = h \times w$ ($[h, w]$ being the size of the feature map). Next, the 1D flattened features are encoded and decoded by the multi-head self-attention units as given in the work of [93].

3.2.3 Multi-modal machine vision tasks

The machine vision (MV) tasks in this category include vision-language tasks like visual question-answering (VQA) [4], visual commonsense-reasoning (VSR) [114], crossmodal retrieval [58] and image-captioning [94]. There is a body of work for these areas within the scope of this paper, and the notable works have been mentioned here. In their work [85], the authors propose VL-BERT [85], one such technique for learning features which can be generalized to multi-modal MV downstream tasks like VSR and VQA. This technique involves aligning both visual as well as linguistic cues

in order for learning compositely and effectively. For this, [85] uses the BERT (Bidirectional encoder representations from transformers) [17] architecture, and feeds it the features obtained from both visual and language domains. The language-features are the tokens in the input text sequences and the visual-features are the ROIs obtained from the input image by using a standard faster R-CNN model [77]. Their performance on various multi-modal MV tasks shows the advantage of the proposed technique over conventional 'language only' pre-training as done in the BERT [17].

3.2.4 Video understanding

Videos which are audiovisual data are abundantly found. In spite of this, the contemporary techniques tend to learn from short videos (up to few seconds) allowing them to interpret usually short-range relationships [93,37]. Long-range relationship learning is needed in different uni-modal and multi-modal MV tasks like activity recognition [44,9,25,80,98]. In this section, we highlight some recent techniques from the CNN transformer pipeline domain which seek to address this issue better than transformer networks.

In their work [118], the authors study the problem of dense-video captioning with transformers. This requires producing language data for every event occurring in the video. The earlier techniques used for the same usually proceed sequentially: event-detection followed by caption-generation inside distinct sub-blocks. The authors of [118] propose a unified transformer architecture which learns one model for tackling both the aforementioned tasks jointly. Thus the proposed technique combines both the multi-modal MV tasks of event-detection and caption-generation. In the first stage, a video-encoder has been used for obtain frame wise features, which is followed by 2 decoder units which propose relevant events and related captions. As a matter of fact, [118] is the first technique for dense-video captioning without using recurrent models. It uses self-attention based encoder which is fed CNN output features. Experimentation on ActivityNet Captions [46] and YouCookII [117] datasets reported valuable improvement over earlier RNN and double-staged techniques.

In their work [59], the authors have noted in their work that in the multi-modal MV task learning techniques like VideoBERT [86] and ViLBERT [68] the language-processing part is generally kept fixed for a pre-trained model like BERT [17] for reducing the training complexity. As an alternative and also as a first, they have proposed PEMT, a multi-modal bidirectional transformer which can learn end-to-end audio-visual video data. In their model, short-term dependencies are first learnt using CNNs, and this is followed by a long-term dependency learning unit. The technique uses CNN features learned during its training for selection of negative samples which are similar to positive samples. The results obtained show that the concept has good implications on multi-modal task model performance.

Traditionally, CNN-based techniques for video classification usually performed 3D spatio-temporal manipulation on relatively small intervals for video understanding. In their work [7], the authors have proposed the video transformer network (VTN) which first obtains frame features from a 2D CNN then applies a transformer encoder for learning temporal relationships. There are 2 advantages of using transformer encoder for the spatial features: (i) whole video is processed in a single pass,

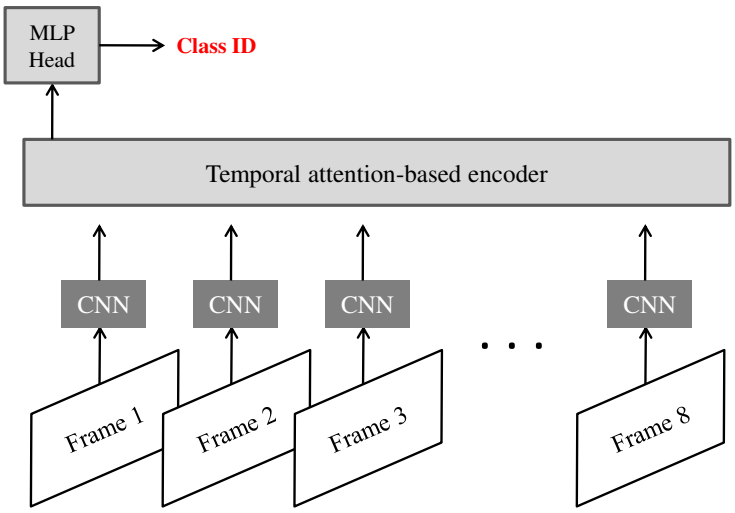

Fig. 6: Video transformer network (VTN) architecture [7]

and (ii) training and efficiency are improved considerably by avoiding 3D convolution which is expensive.

These feats make VTN suitable for learning from long videos in which inter-entity interactions are spread length-wise. The experiments of the authors on the Kinetics-400 dataset [44] with various CNN and non-CNN backbones e.g. ResNet [34], ViT [20] and DeiT [90], show good performance. 6 shows the overall schematic of the proposed model.

In the next sub-section, we turn to third category of techniques of attention mechanisms and deep learning in machine vision, i.e. hybrid transformers.

3.3 Hybrid transformers

Transformers used to be exclusively attention based networks. However, some recent works have introduced two new variants i.e. convolutional vision transformers (CvTs) and hybrid CNN-transformer models. These variants are discussed below.

3.3.1 Convolutional vision transformers

In natural language processing (NLP) and speech recognition (SR), convolutional operations were used for modification of the transformer unit, either by changing the multi-head attention blocks with convolutional layers [102], or by adding more parallel convolutional layers [104] or more sequential convolutional layers [28], in order to capture local dependencies. Earlier research [99] proposed propagation of the attention maps to following layers by residual connections being transformed by convolutional operations.

Convolutional vision transformers (CvTs) [103] improve the vision transformer (ViT) both in terms of performance and efficiency with the introduction of convolutions into ViT for yielding the best of both architectures. This has been achieved through 2 important modifications. First, a range of transformers with a novel convolutional token-embedding and second, a convolutional transformer unit giving convolutional projections. Thus they propose introduction of convolutional operations to 2 primary parts of the ViT viz., first, replacement of the 'linear projection' used for every position in the attention mechanism with their novel 'convolutional projection', and second, use of their hierarchical multistage architecture for enabling variable resolution of two-dimensional reshaped tokens just like CNNs. These fundamental changes have introduced desirable properties of CNNs to ViTs i.e., shift-, scale-, and distortion-invariance, while at the same time have maintained the merits of transformers i.e. global context, dynamic attention, and higher level of generalization. The authors validate CvT through extensive experimentation showing that their technique achieves state of the art performance as compared to other ViTs and ResNets on the ImageNet-1k dataset, with lesser parameters and lesser FLOPs. Also, the performance gains stay when CvT is pre-trained on larger datasets like ImageNet-22k [16] and is subsequently fine-tuned for downstream tasks. Pre-training on ImageNet-22k leads to top-1 accuracy of 87.7% for the ImageNet-1k validation set. Lastly, their results demonstrate that positional encoding which is an important component in existing ViTs, can be suitably removed in CvT thus simplifying its architecture for higher resolution MV tasks.

3.3.2 Hybrid CNN-transformer models

A wide range of recent developments in handcrafted neural network models for machine vision tasks have asserted the important need for exploration of hybrid models which consist of diverse building blocks. At the same time, neural network model searching techniques are surging with expectations of reduction in human effort. In evidence brought out by some works [19,84,6] it is stated that hybrids of convolutional neural networks (CNNs) and transformers can perform better that both pure CNNs and pure transformers. In spite of this, the question that whether neural architecture search (NAS) methods can handle different search spaces with different candidates like CNNs and transformers, effectively and efficiently, leads to an open research area. In their work [60], the authors propose the 'block-wisely self-supervised neural architecture search' (BossNAS) which is an unsupervised NAS technique which addresses the issue of inaccurate model rating due to large weight-sharing space and supervision with bias as undertaken in earlier techniques. Going into specifics, they factorize the search-space into smaller blocks and also utilize a new self-supervision based training technique called 'ensemble bootstrapping', for training every block individually prior to search. Also, they propose a search-space called HyTra which is like a hybrid search-space fabric of CNNs and transformers. The fabric like search-space consists of model architectures similar to the common ViTs [19,91,12], CNNs [35,39] and hybrid CNN-transformers [84] at various scales. Over the same difficult search-space, their searched hybrid model viz. BossNet-T yields 82.2% accuracy for ImageNet, going beyond EfficientNet by a

margin of 2.1% with similar computation time. Also, they report that their technique achieves better model rating accuracy on the MBConv search-space for ImageNet and on NATS-Bench size search-space for CIFAR-100 than the state of the art NAS techniques. The code and the pre-trained models are available online at https://github.com/changlin31/BossNAS.

In the next section, we discuss the major research algorithms, issues and trends in techniques of attention and deep learning in machine vision.

4 Major research algorithms, issues and trends

In the field of machine vision (MV), recently attention based mechanisms are generating a lot of interest. Pure attention based architectures/models are slowly and steadily proving worthy of loosening the grip of deep learning over MV as interesting and efficient attention based models continue to be built. However, pure attention based models come with their own set of issues. They are quite 'data-hungry' as they require huge amounts of data to pre-train before being able to be applied to MV downstream tasks after fine-tuning. As an example, vision transformers have to be pre-trained on the JFT dataset [2] which consists of 300 million images, and subsequently have to be fine-tuned on ImageNet-1K [15] before they can be used for MV tasks like image classification/retrieval. Also, the training times are exceedingly long for pre-training in transformers. Hence, reducing the 'hunger/appetite' of transformers is an open research area. Also, reducing the training time of transformers by using efficient architectures and training techniques is also an open research area. Reducing the computational load/resources for training of vision based transformers is also an open research area besides finding novel ways to port them to limited hardware/resource (portable) platforms available in the industry. A very large body of research work is present on deep learning and CNN based architectures and transformers can benefit from the same, as CNN based models have taken a foothold in MV. The industrial footprint of deep learning and CNN based models is also large. Attention based models can benefit from the work done and industrial footprint of deep learning based models. Some works [19,84,6] state that hybrids of convolutional neural networks (CNNs) and transformers can perform better that both pure CNNs and pure transformers.

Currently, the algorithms applicable to transformers benefitting from deep learning and CNN architecture are present in three main categories as discussed earlier. The first category being attention-based CNNs. The algorithms in this category aim to augment the performance of classical CNN architectures by plugging into them attention-based components/units in order to refine the features as and when they are used. Attention based CNN plugins like CBAM have been used successfully in various CNNs models/architectures to boost their performance at relatively small computational time overhead. In spite of this, the amount of attention available in this category is limited and the CNNs use the attention based mechanisms sparingly. Deeper integration and merging of attention based mechanisms and CNNs are required before outstanding and record breaking performances can be achieved. Coming to the second category of CNN transformer pipelines which has also been discussed, the

pipeline is just like the earlier hybrid two-stage classifiers wherein a feature map generated by a 'teacher' CNN is fed to a waiting 'student' classifier which operates on this feature map. In this two-stage model, it is safe to say that the performance of the second-stage model depends on the image/video interpretation capability/capacity of the CNN. As such the architecture/design of the first-stage CNN is in question regarding its design-based efficacy at efficiently interpreting the image/video data. And it is known that there are currently a large number of CNN architectures available and making the correct choice is an open research field. Coming to the third category of Hybrid CNN-transformers, the merging of these two different techniques is a difficult one. Network architecture search (NAS) has been used to search through hybrid CNN-transformer search-space fabric. However, given its exhaustive nature requiring large computational resources and careful fabric design, the optimization of the same is also an open research area. In spite of the limitations and issues mentioned above, attention based mechanisms like vision transformers (ViTs) are considered having potential to impact the MV research and industrial body in the future. Combined with the power and experience of deep learning, the merger of the two techniques can prove to be revolutionary for both the existing and as well as the upcoming machine vision (MV) tasks/applications, as new, larger and more efficient computational hardware and software continue to be developed.

5 Conclusion

In this paper, the merger of attention based mechanisms and deep learning for various machine vision (MV) tasks/applications has been discussed. In the beginning of the paper, various types of attention mechanism were briefly discussed. Next, various attention based architectures were discussed. This was followed by discussing various categories of combinations of attention mechanisms and deep learning techniques for machine vision (MV). The various architectures and their associated machine vision tasks/applications were discussed. Afterwards, major research algorithms, issues and trends within the scope of the paper were discussed. By using 110+ papers as research reference in this survey, the readers of this paper are expected to form a knowledge-base and get a head-start in the area of combinational techniques of attention based mechanisms and deep learning for machine vision.

Conflict of interest

The authors declare no conflict of interest.

References

1. URL https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
2. Revisiting the unreasonable effectiveness of data. URL https://ai.googleblog.com/2017/07/revisiting-unreasonable-effectiveness.html
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018). DOI 10.1109/CVPR.2018.00636

4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2425–2433 (2015). DOI 10.1109/ICCV.2015.279
5. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention (2015)
6. Bello, I.: Lambdanetworks: Modeling long-range interactions without attention. In: International Conference on Learning Representations (2021). URL https://openreview.net/forum?id=xTJEN-ggl1b
7. Berg, A., O'Connor, M., Cruz, M.T.: Keyword transformer: A self-attention model for keyword spotting (2021)
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: A. Vedaldi, H. Bischof, T. Brox, J.M. Frahm (eds.) Computer Vision - ECCV 2020, pp. 213–229. Springer International Publishing, Cham (2020)
9. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset (2019)
10. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer (2020)
11. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306 (2017). DOI 10.1109/CVPR.2017.667
12. Chu, X., Tian, Z., Zhang, B., Wang, X., Wei, X., Xia, H., Shen, C.: Conditional positional encodings for vision transformers (2021)
13. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience **3**(3), 201–215 (2002)
14. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 3488–3493 (2016). DOI 10.1109/ICPR.2016.7900174
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). DOI 10.1109/CVPR.2009.5206848
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). DOI 10.1109/CVPR.2009.5206848
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). DOI 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423
18. Doersch, C., Gupta, A., Zisserman, A.: Crosstransformers: spatially-aware few-shot transfer (2021)
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Mindere, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)
21. Er, M.J., Zhang, Y., Wang, N., Pratama, M.: Attention pooling-based convolutional neural network for sentence modelling. Information Sciences **373**, 388–403 (2016). DOI 10.1016/j.ins.2016.08.084. URL https://www.sciencedirect.com/science/article/pii/S0020025516306673
22. Escalera, S., Baró, X., Gonzàlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: L. Agapito, M.M. Bronstein, C. Rother (eds.) Computer Vision - ECCV 2014 Workshops, pp. 459–473. Springer International Publishing, Cham (2015)
23. Everingham, M., Williams, C.K.: The pascal visual object classes challenge 2007 (voc2007) results
24. Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., Schlaefer, A.: Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. IEEE Transactions on Biomedical Engineering **67**(2), 495–503 (2020). DOI 10.1109/TBME.2019.2915839
25. Ging, S., Zolfaghari, M., Pirsiavash, H., Brox, T.: Coot: Cooperative hierarchical transformer for video-text representation learning (2020)

26. Girdhar, R., João Carreira, J., Doersch, C., Zisserman, A.: Video action transformer network. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 244–253 (2019). DOI 10.1109/CVPR.2019.00033
27. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
28. Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R.: Conformer: Convolution-augmented Transformer for Speech Recognition. In: Proc. Interspeech 2020, pp. 5036–5040 (2020). DOI 10.21437/Interspeech.2020-3015. URL 'http://dx.doi.org/10.21437/Interspeech.2020-3015
29. Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval **7**(2), 87–93 (2018). DOI 10.1007/s13735-017-0141-z
30. Hafiz, A.M., Bhat, G.M.: A survey of deep learning techniques for medical diagnosis. In: M. Tuba, S. Akashe, A. Joshi (eds.) Information and Communication Technology for Sustainable Development, pp. 161–170. Springer Singapore, Singapore (2020)
31. Hafiz, A.M., Bhat, G.M.: A survey on instance segmentation: state of the art. International Journal of Multimedia Information Retrieval **9**, 171–189 (2020). DOI 10.1007/s13735-020-00195-x
32. Hang, R., Li, Z., Liu, Q., Ghamisi, P., Bhattacharyya, S.S.: Hyperspectral image classification with attention-aided cnns. IEEE Transactions on Geoscience and Remote Sensing **59**(3), 2281–2293 (2021). DOI 10.1109/TGRS.2020.3007921
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE Computer Society, Los Alamitos, CA, USA (2016). DOI 10.1109/CVPR.2016.90. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90
34. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
36. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
37. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997). DOI 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735
38. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(8), 2011–2023 (2020). DOI 10.1109/TPAMI.2019.2913372
39. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Trans. Pattern Anal. Mach. Intell. **42**(8), 2011–2023 (2020). DOI 10.1109/TPAMI.2019.2913372. URL https://doi.org/10.1109/TPAMI.2019.2913372
40. Huang, J., Zhou, W., Li, H., Li, W.: Attention-based 3d-cnns for large-vocabulary sign language recognition. IEEE Transactions on Circuits and Systems for Video Technology **29**(9), 2822–2832 (2019). DOI 10.1109/TCSVT.2018.2870740
41. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 262–270 (2015). DOI 10.1109/ICCV.2015.38
42. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 262–270 (2015). DOI 10.1109/ICCV.2015.38
43. Jetley, S., Murray, N., Vig, E.: End-to-end saliency mapping via probability distribution prediction. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5753–5761 (2016). DOI 10.1109/CVPR.2016.620
44. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017)
45. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey (2021)
46. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 706–715 (2017). DOI 10.1109/ICCV.2017.83
47. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009)
48. Kruthiventi, S.S.S., Ayush, K., Babu, R.V.: Deepfix: A fully convolutional neural network for predicting human eye fixations. IEEE Transactions on Image Processing **26**(9), 4446–4456 (2017). DOI 10.1109/TIP.2017.2710620

49. Kruthiventi, S.S.S., Gudisa, V., Dholakiya, J.H., Babu, R.V.: Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5781–5790 (2016). DOI 10.1109/CVPR.2016.623
50. Kumar, M., Weissenborn, D., Kalchbrenner, N.: Colorization transformer (2021)
51. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet (2015)
52. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Deepgaze ii: Reading fixations from deep features trained on object recognition (2016)
53. Larochelle, H., Hinton, G.: Learning to combine foveal glimpses with a third-order boltzmann machine. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1, NIPS'10, pp. 1243–1251. Curran Associates Inc., Red Hook, NY, USA (2010)
54. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
55. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998). DOI 10.1109/5.726791
56. LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, pp. 253–256 (2010). DOI 10.1109/ISCAS.2010.5537907
57. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-Supervised Nets. In: G. Lebanon, S.V.N. Vishwanathan (eds.) Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 38, pp. 562–570. PMLR, San Diego, California, USA (2015). URL http://proceedings.mlr.press/v38/lee15a.html
58. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
59. Lee, S., Yu, Y., Kim, G., Breuel, T., Kautz, J., Song, Y.: Parameter efficient multimodal transformers for video representation learning (2020)
60. Li, C., Tang, T., Wang, G., Peng, J., Wang, B., Liang, X., Chang, X.: Bossnas: Exploring hybrid cnn-transformers with block-wisely self-supervised neural architecture search (2021)
61. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5455–5463 (2015). DOI 10.1109/CVPR.2015.7299184
62. Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z.: Attention mechanism-based cnn for facial expression recognition. Neurocomputing **411**, 340–350 (2020). DOI https://doi.org/10.1016/j.neucom.2020.06.014. URL https://www.sciencedirect.com/science/article/pii/S0925231220309838
63. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10563–10572 (2019). DOI 10.1109/CVPR.2019.01082
64. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
65. Li, Y., Zeng, J., Shan, S., Chen, X.: Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Transactions on Image Processing **28**(5), 2439–2450 (2019). DOI 10.1109/TIP.2018.2886767
66. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds.) Computer Vision - ECCV 2014, pp. 740–755. Springer International Publishing, Cham (2014)
67. Liu, N., Han, J., Liu, T., Li, X.: Learning to predict eye fixations via multiresolution convolutional neural networks. IEEE Transactions on Neural Networks and Learning Systems **29**(2), 392–404 (2018). DOI 10.1109/TNNLS.2016.2628878
68. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks (2019)
69. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101 (2010). DOI 10.1109/CVPRW.2010.5543262
70. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205 (1998). DOI 10.1109/AFGR.1998.670949

71. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing **10**(1), 18–31 (2019). DOI 10.1109/TAFFC.2017.2740923
72. Pan, J., Sayrol, E., Giro-I-Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 598–606 (2016). DOI 10.1109/CVPR.2016.71
73. Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P.: Learning to refine object segments. In: B. Leibe, J. Matas, N. Sebe, M. Welling (eds.) Computer Vision - ECCV 2016, pp. 75–91. Springer International Publishing, Cham (2016)
74. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollar, P.: Designing network design spaces. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10425–10433 (2020). DOI 10.1109/CVPR42600.2020.01044
75. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Chen, M., Child, R., Misra, V., Mishkin, P., Krueger, G., Agarwal, S., et al.: Dall· e: Creating images from text. OpenAI blog. https://openai. com/blog/dall-e (2021)
76. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15, pp. 91–99. MIT Press, Cambridge, MA, USA (2015)
77. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017). DOI 10.1109/TPAMI.2016.2577031
78. Rensink, R.A.: The dynamic representation of scenes. Visual Cognition **7**(1-3), 17–42 (2000). DOI 10.1080/135062800394667. URL https://doi.org/10.1080/135062800394667
79. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017). DOI 10.1109/ICCV.2017.74
80. Seong, H., Hyun, J., Kim, E.: Video multitask transformer network. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 1553–1561 (2019). DOI 10.1109/ICCVW.2019.00194
81. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. In: International Conference on Learning Representations (ICLR) Workshop (2016). URL https://arxiv.org/abs/1511.04119
82. Shrestha, A., Mahmood, A.: Review of deep learning algorithms and architectures. IEEE Access **7**, 53040–53065 (2019). DOI 10.1109/ACCESS.2019.2912200
83. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
84. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition (2021)
85. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (2020). URL https://openreview.net/forum?id=SygXPaEYvH
86. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7463–7472 (2019). DOI 10.1109/ICCV.2019.00756
87. Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5100–5111. Association for Computational Linguistics, Hong Kong, China (2019). DOI 10.18653/v1/D19-1514. URL https://www.aclweb.org/anthology/D19-1514
88. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: K. Chaudhuri, R. Salakhutdinov (eds.) Proceedings of the 36th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 97, pp. 6105–6114. PMLR (2019). URL http://proceedings.mlr.press/v97/tan19a.html
89. Tian, C., Xu, Y., Li, Z., Zuo, W., Fei, L., Liu, H.: Attention-guided cnn for image denoising. Neural Networks **124**, 117–129 (2020). DOI https://doi.org/10.1016/j.neunet.2019.12.024. URL https://www.sciencedirect.com/science/article/pii/S0893608019304241
90. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention (2021)

91. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention (2021)
92. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
93. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, u., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pp. 6000–6010. Curran Associates Inc., Red Hook, NY, USA (2017)
94. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164 (2015). DOI 10.1109/CVPR.2015.7298935
95. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450–6458 (2017). DOI 10.1109/CVPR.2017.683
96. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Transactions on Image Processing **27**(5), 2368–2378 (2018). DOI 10.1109/TIP.2017.2787612
97. Wang, X., Yeshwanth, C., Niebner, M.: Sceneformer: Indoor scene generation with transformers (2021)
98. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers (2021)
99. Wang, Y., Yang, Y., Bai, J., Zhang, M., Bai, J., Yu, J., Zhang, C., Huang, G., Tong, Y.: Evolving attention with residual convolutions (2021)
100. Woo, S., Hwang, S., Kweon, I.S.: Stairnet: Top-down semantic aggregation for accurate one shot detection. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1093–1102 (2018). DOI 10.1109/WACV.2018.00125
101. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds.) Computer Vision - ECCV 2018, pp. 3–19. Springer International Publishing, Cham (2018)
102. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions (2019)
103. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers (2021)
104. Wu, Z., Liu, Z., Lin, J., Lin, Y., Han, S.: Lite transformer with long-short range attention (2020)
105. Xie, S., Tu, Z.: Holistically-nested edge detection. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1395–1403 (2015). DOI 10.1109/ICCV.2015.164
106. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: F. Bach, D. Blei (eds.) Proceedings of the 32nd International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 37, pp. 2048–2057. PMLR, Lille, France (2015). URL http://proceedings.mlr.press/v37/xuc15.html
107. Xu, M., Li, C., Liu, Y., Deng, X., Lu, J.: A subjective visual quality assessment method of panoramic videos. In: 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 517–522 (2017). DOI 10.1109/ICME.2017.8019351
108. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013). DOI 10.1109/CVPR.2013.407
109. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5790–5799 (2020). DOI 10.1109/CVPR42600.2020.00583
110. Yao, X., Han, J., Zhang, D., Nie, F.: Revisiting co-saliency detection: A novel approach based on two-stage multi-view spectral rotation co-clustering. IEEE Transactions on Image Processing **26**(7), 3196–3209 (2017). DOI 10.1109/TIP.2017.2694222
111. Ye, H.J., Hu, H., Zhan, D.C., Sha, F.: Few-shot learning via embedding adaptation with set-to-set functions. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8805–8814 (2020). DOI 10.1109/CVPR42600.2020.00883
112. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10494–10503 (2019). DOI 10.1109/CVPR.2019.01075

113. Yu, Y., Choi, J., Kim, Y., Yoo, K., Lee, S.H., Kim, G.: Supervising neural attention models for video captioning by human gaze data. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6119–6127 (2017). DOI 10.1109/CVPR.2017.648
114. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6713–6724 (2019). DOI 10.1109/CVPR.2019.00688
115. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikäinen, M.: Facial expression recognition from near-infrared videos. Image and Vision Computing **29**(9), 607–619 (2011). DOI https://doi.org/10.1016/j.imavis.2011.07.002. URL https://www.sciencedirect.com/science/article/pii/S0262885611000515
116. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1265–1274 (2015). DOI 10.1109/CVPR.2015.7298731
117. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (2018). URL https://ojs.aaai.org/index.php/AAAI/article/view/12342
118. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8739–8748 (2018). DOI 10.1109/CVPR.2018.00911
119. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection (2021)