Whitening Transformation inspired Self-Attention for Powerline Element Detection

Emmanouil Patsiouras
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
emmanoup@csd.auth.gr

Vasileios Mygdalis
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
mygdalisv@csd.auth.gr

Ioannis Pitas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece
pitas@csd.auth.gr

***Abstract*—Powerline inspection operations involve capturing and inspecting visual footage of powerline elements from elevated positions above and around the powerline and are currently performed with the help of helicopters and/or Unmanned Aerial Vehicles (UAVs). Current technological advances in the areas of robotics and machine learning are towards enabling fully autonomous operations. To this end, one of the tasks to be addressed is the robust, precise and fast powerline object detection problem. Recently introduced Transformer-based object detection methods demonstrate time and accuracy advances with respect to previous works. In this work, we present an enhanced Transformer-based architecture that further improves the state-of-the-art by incorporating a content-specific object query generator and by substituting the original attention operation with a whitening-inspired transformation at certain stages of the architecture. We evaluate our method in a recently captured powerline detection dataset and we show that our novel contributions offer a significant boost regarding detection accuracy.**

***Keywords*—deep learning, object detection, attention models, powerline inspection**

I. Introduction

Powerline inspection operations involve gathering detailed video footage from elevated positions close and around powerline elements [1]. To this end, we turn our focus on the object detection task, where the goal is to visually localize, identify and monitor specific elements and components in high-voltage transmission lines, such as towers, insulators and dumpers. This task, however, poses challenges that may not be present in standard benchmark datasets such as MS-COCO [2]. More specifically, the successful detection of powerline elements requires high precision in both small and big powerline elements appearing on the same image frame (e.g. towers and insulators), robustness in difficult illumination conditions (sometimes even against the sun), and the ability to discriminate visually similar and small elements on the powerline against an indistinguishable background.

To tackle the aforementioned task we examine employing a state-of-the-art object detection method, namely the Detection with Transformers (DETR) [3] detector. DETR's major benefit is attributed to the attention [4] mechanism encountered in the Transformers [5] architecture. Given N input data sequences (e.g. image patches), attention mechanisms aggregate information from the entire sequence (i.e., the whole image) to each sequence element. This aggregated information explicitly models all pairwise interactions between all elements in the sequences thus being able to use the whole sequence as context.

DETR, however, despite its fascinating design and formidable performance comes with its own shortcomings. One such is its rather poor ability to detect small objects such as in our experimental application scenario, mainly attributed to the fact that it ends up utilizing low-resolution feature maps, hence small depicted objects in the original image become indistinguishable in the intermediate representation stages. This issue could be addressed by utilizing higher resolution feature maps but simultaneously leading to unacceptable computational complexities imposed by the attention operation when dealing with large sequences [6]. Another possible drawback could derive from the fact that DETR uses an arbitrary number of randomly initialized learnable object queries (explained in following sections). This random initialization could possibly tamper with the framework's performance as much of the network's domain knowledge is stored in these object queries.

In this paper we propose two novel contributions in order to address the aforementioned problems and offer a boost in detection performance. Firstly, we propose a content-specific object query generator, in the form of a very shallow convolutional neural network (CNN) which produces object queries that hold information related to the images in the dataset. Secondly, we present a whitening transformation-inspired [7] self-attention definition and we use this re-formulation to substitute the original attention definition only at certain stages of the whole framework. This has the attribute of regularizing illumination change variations and at the same time, allowing the model to capture semantic dependencies between the image depicted objects.

The rest of the paper is structured as follows. Section II overviews the object detection task and reviews the key points of attention and whitening transformation. In Section III we introduce and analyze our novel contributions. In Section IV we present and discuss the experiments conducted on our powerline element dataset. Finally, Section V concludes and summarizes our work.

II. Revisiting Object Detection and Attention

The task of object detection [8] consists of classifying and localizing every object of interest depicted in an image frame, in the form of rectangular Regions of Interest (ROI). In this work, we focus on the application scenario of linear infrastructure inspection, which requires the precise localization of powerline elements and components on the power transmission lines. The most common way of representing these objects is through predicting a set of bounding boxes associated with category labels. Most of the popular Deep Learning (DL) Convolutional Neural Network (CNN)-based object detectors [9] treat this task as a combination of classification and bounding box regression and depending on whether or not they rely on region proposals they fall into one of two categories, namely two-stage [10], [11] or single-stage detectors [12], [13], respectively. Finally, a Non-Maximum Suppression (NMS) step is performed to eliminate multiple detections of the same object in overlapping regions.

However, DETR addresses the object detection as a direct set prediction problem and completely eliminates the need of any geometric priors resulting in the first fully differentiable end-to-end object detector. One major novelty introduced in DETR was the incorporation of Transformer [5] neural network blocks in conjunction with a CNN model, in an object detection pipeline. Transformer's dominance over a range of different tasks [14], [15], [16] derives from the attention operation and in the following subsections we review in details this mechanism.

A. Attention

In general, Transformers constitute an attention mechanism that receive two sequence of elements as inputs (e.g. texts or image feature maps), and updates the elements of one of them by aggregating information from every element of the other.

Let $\mathbf{X} \in \mathbb{R}^{N\times n}$ and $\mathbf{Y} \in \mathbb{R}^{M\times n}$ be two input matrices consisting of N and M elements respectively of n dimension each. Based on $\mathbf{X}, \mathbf{Y}$ the following three matrices can be created: i) a query matrix $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q + \mathbf{1}_{N\times 1}\mathbf{b}_Q$, ii) a key matrix $\mathbf{K} = \mathbf{Y}\mathbf{W}_K + \mathbf{1}_{M\times 1}\mathbf{b}_K$, and iii) a value matrix $\mathbf{V} = \mathbf{Y}\mathbf{W}_V + \mathbf{1}_{M\times 1}\mathbf{b}_V$. $\mathbf{W}_Q \in \mathbb{R}^{n\times n_q}, \mathbf{W}_K \in \mathbb{R}^{n\times n_k}$ and $\mathbf{W}_V \in \mathbb{R}^{n\times n_v}$ are linear transformation matrices applied on the input matrices and $\mathbf{b}_Q \in \mathbb{R}^{n_q}, \mathbf{b}_K \in \mathbb{R}^{n_k}, \mathbf{b}_V \in \mathbb{R}^{n_v}$ are their respective biases. For simplicity we can consider $n_q = n_k = n_v = n$. Based on the above, the operation of attention is now defined as:

$$\mathbf{A} = \overbrace{\sigma(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{n}})}^{\mathbf{S}}\mathbf{V}, \tag{1}$$

where $\sigma(\cdot)$ is the softmax operator applied on every row of the matrix $\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{n}}$. This type of attention is also called cross-attention. Self-attention is defined when $\mathbf{Y}$ coinsides with $\mathbf{X}$ (i.e. $\mathbf{Y} = \mathbf{X}$) or vice versa. Self-attention is the basic building block of a Transformer model and allows it to associate each element in an input sequence to every other element in the same sequence.

The above definition describes a naive attention computation. In practice, however, Transformers usually leverage a multi-headed attention mechanism. In this case we have N_h number of attention heads and we split the $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ matrices into N_h matrices of dimensions $n \times \frac{n}{N_h}$ (n should be divisible by N_h). The attention for every single head, $\mathbf{A}_h, h = 1, \ldots, N_h$ is defined as:

$$\mathbf{A}_h = \overbrace{\sigma(\frac{\mathbf{Q}_h\mathbf{K}_h^T}{\sqrt{n}})}^{\mathbf{S}_h}\mathbf{V}_h, \tag{2}$$

where $\mathbf{Q}_h = \mathbf{X}\mathbf{W}_{Q_h}, \mathbf{K}_h = \mathbf{Y}\mathbf{W}_{K_h}$, and $\mathbf{V}_h = \mathbf{Y}\mathbf{W}_{V_h}$ (biases are omitted for simplicity) for $h = 1, \ldots, N_h$. The overall attention is the concatenation of the attention of every single head and is defined as:

$$\mathbf{A} = (\mathbf{A}_1 \oplus \mathbf{A}_2 \oplus \cdots \oplus \mathbf{A}_h)\mathbf{W}_O, \tag{3}$$

where $\oplus$ represents the concatenation of two 2D matrices along the horizontal dimension and $\mathbf{W}_O \in \mathbb{R}^{n\times n_o}$ is a linear projection matrix. Again we can consider $n_o = n$.

B. Whitening transformation

Given a data matrix $\mathbf{X} \in \mathbb{R}^{N\times n}$ consisting of N $n-$dimensional elements the goal is to apply a linear transformation in order to decorrelate the data dimensions from one another. This can be achieved by using a linear transformation matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ which will transform the data matrix $\mathbf{X}$ such as $\mathbf{X}_w = \mathbf{W}\mathbf{X}$.

In order for $\mathbf{W}$ to be a decorrelating matrix it should satisfy the condition $\mathbf{W}\mathbf{W}^T = \mathbf{\Sigma}^{-1}$, where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$.

III. Proposed Method

As already stated, we employ DETR [3] as a testbed for implementation and performance comparison. DETR consists of a CNN backbone for extracting feature representations from input images and an encoder-decoder Transformer module. The CNN follows a standard procedure by receiving an input image $\mathbf{x} \in \mathbb{R}^{H_0\times W_0\times 3}$, where H_0, W_0 are the height and width of the input image, and produce lower-resolution feature maps $\mathbf{f} \in \mathbb{R}^{H\times W\times C}$, where C is the number of output channels. The Transformer encoder receives these unrolled feature maps $\mathbf{z} \in \mathbb{R}^{HW\times C}$ and calculates self-attention as described by equations (2) and (3). The rest of encoder follows a standard architecture as described in [5]. The decoder is of similar structure, however, its input includes both the output of the encoder and N_q in number $n-$dimensional object queries represented by learnable and random initialized query embeddings $\mathbf{o} \in \mathbb{R}^{N_q\times n}$. Decoder calculates self-attention between the object queries and subsequently performs cross-attention between the transformed object queries and the transformed feature maps from the encoder. Finally, DETR passes the final output of the decoder, i.e. the transformed object queries, to

a feed forward network that predicts for every object query either a detection (bounding box and class) or a no object class. A simplified representation of the overall architecture along with our novel modifications is illustrated in Fig. 1.

We will now describe our two novel contributions for tackling the powerline element detection using the Transformer-based object detector.

A. Whitened self-attention

As mentioned in the introduction Section, our application scenario for testing our detection models is that of detecting elements and components most commonly found in high-voltage transmission lines for inspection purposes. However, this detection task comes with quite a few challenges. Maybe the most prominent challenges include the requirement to simultaneously detect small and big electrical components that appear from short and long distances, diverse viewing angles and with various illumination conditions. Further challenges could be imposed when the objects of interest are clutched against an indistinguishable background. In order to address these challenges, we introduce the whitening self-attention variation, which eliminates some of the variability of the feature space-related mostly to illumination changes. This in turn provides the model for the opportunity to work in a more standardized input space, thus the learning problem can focus more on addressing the semantic challenges, related to small/big item appearances.

In Section II we described the original attention definition and by examining equation (1) (for simplicity) we can determine that the most important term is the calculation of the similarity matrix $\mathbf{S} = \sigma(\mathbf{Q}\mathbf{K}^T)$ which in the case of self-attention can be rewritten as:

$$\mathbf{S} = \sigma(\mathbf{X}\mathbf{W}_Q(\mathbf{X}\mathbf{W}_K)^T) = \sigma(\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{X}^T). \quad (4)$$

By requiring that the auxiliary matrices $\mathbf{W}_Q = \mathbf{W}_K = \mathbf{W}$ such that $\mathbf{W}$ is a decorrelating matrix, equation (4) can be re-formulated as:

$$\mathbf{S} = \sigma(\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}^T), \quad (5)$$

where $\boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^T$ is the covariance matrix of $\mathbf{X}$. The covariance matrix $\boldsymbol{\Sigma}$ can be computed as follows:

$$\begin{aligned}\boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})^T(\mathbf{x}_i - \overline{\mathbf{x}}) &= \mathbf{X}^T\overbrace{\left(\frac{1}{N}\mathbf{I} - \frac{1}{N^2}\mathbf{1}\mathbf{1}^T\right)}^{\mathbf{L}}\mathbf{X} \\ &= \mathbf{X}^T\mathbf{L}\mathbf{X}.\end{aligned} \quad (6)$$

The next step is to calculate the inversion of this matrix $\boldsymbol{\Sigma}^{-1}$. Based on all of the above we re-formulate the self-attention operation as:

$$\mathbf{A} = \overbrace{\sigma\left(\frac{\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{X}^T}{\sqrt{n}}\right)}^{\mathbf{S}}\mathbf{X}\mathbf{W}_V = \overbrace{\sigma\left(\frac{\mathbf{L}^{-1}}{\sqrt{n}}\right)}^{\mathbf{S}}\mathbf{X}\mathbf{W}_V. \quad (7)$$

Based of equation (7) we observe that for computing the similarity matrix $\mathbf{S}$ of the self-attention operation we require only the matrix $\mathbf{L}$, the softmax inverted version of which can be precomputed before initiating any training phase. However, in order to avoid singularities arising from matrix invertion, we introduce a regularization parameter that increases the rank of the covariance matrix $\boldsymbol{\Sigma}$ by adding tiny values to the diagonal elements:

$$\boldsymbol{\Sigma} = \mathbf{X}^T\mathbf{L}\mathbf{X} + r\mathbf{I}, \quad (8)$$

where r is typically set to very low values (a value of $r = 10^{-3}$ was employed in our experimental study). For all of our experiments that include our novel self-attention definition we use (8) for computing equation (7).

In all of the above we eliminated the $\mathbf{W}_Q$ and $\mathbf{W}_K$ learnable matrices with the additional cost of having to invert the quantity inside the parenthesis. The above definition is only recommended to be employed at the first layer of the Transformer encoder whereas for the rest of the encoder we use the standard attention definition. This is done in order to both eliminate the inversion operation in multiple layers, but also to avoid one additional particularity. Whitening in the first layer might decorrelate variations of illumination conditions, however, applying it to later stages might also eliminate the semantic information that the model is trying to learn. Our approach replaces the $\mathbf{W}_Q$ and $\mathbf{W}_K$ matrices with appropriate non-learnable ones that perform the whitening transform instead of a random transformation. This operates as adapting the input domain of the self-attention operation to a latent space domain, where only query and key image features have been normalized according to their own distribution. At this point, we should note that applying our novel self-attention, only at the first layer of a Transformer encoder, is not the same as simply applying whitening or some standard layer normalization to the input data $\mathbf{X}$ before feeding them to a standard self-attention architecture because this would imply whitening of the value matrix $\mathbf{V}$ as well. In our experimental study we conduct such comparisons to prove the superiority of our approach.

B. Content-specific object query generator

As explained in previous sections, DETR uses a set of learnable object queries that will eventually be transformed to a set of detections. However, the random initialization of those object queries could impose certain limitations to the whole framework's performance. To this end, we propose the employment of a content-specific object queries generator whose job is to produce better initializations directly related to the image features. Our object queries generator is implemented in the form of a very shallow 2-layer CNN that receives image feature maps, as produced by the CNN backbone, and generates content-related start points for the object queries. Specifically, the feature maps $\mathbf{f} \in \mathbb{R}^{H\times W\times C}$ initially pass through a first convolutional layer with C_1 number of channels and 3×3 kernel size with padding, followed by a batch normalization layer [17], a ReLU activation

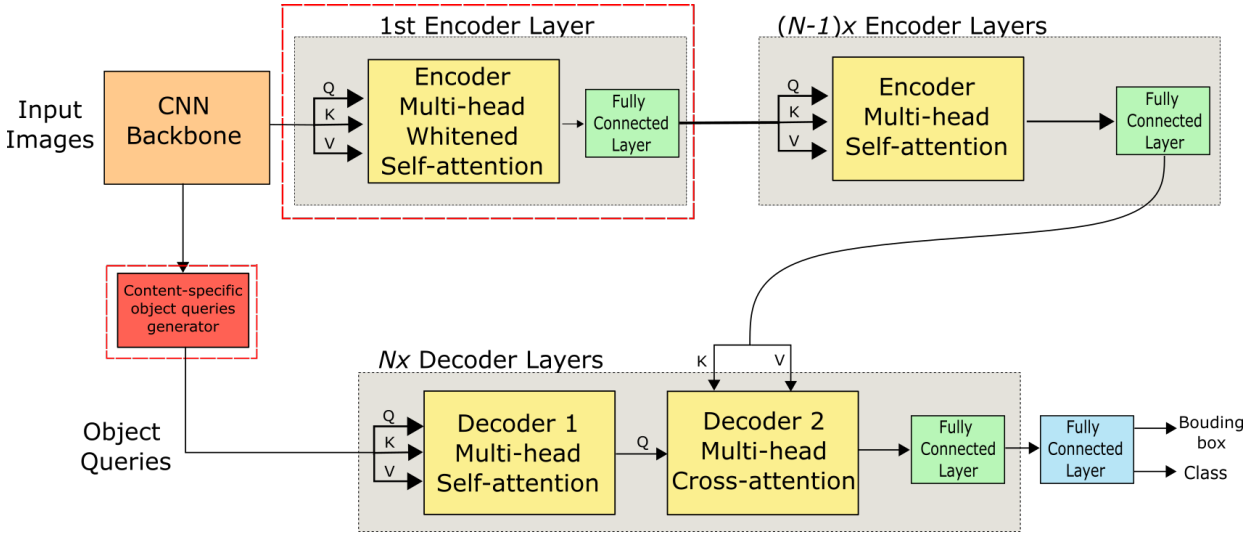

Fig. 1. A simplified illustration of DETR's architecture utilizing Transformer blocks. In red squares we highlight our novel contributions, i.e., a) the re-formulation of the self-attention operation in the first Transformer encoder layer, and b) the object query generator. The object query generator specifically is implemented as a 2-layer CNN receiving image features from the CNN backbone and producing content-specific object queries.

function and a dropout layer [18]. The second layer is of similar structure with C_2 number of channels without applying padding. The output of this query generator is a new set of feature maps $\mathbf{f}' \in \mathbb{R}^{H' \times W' \times C_2}$. These flattened feature maps $\mathbf{p} \in \mathbb{R}^{H'W' \times C_2}$ constitute the initializations for the object queries. It can be noticed that our number of object queries, i.e. $H'W'$, is not so arbitrary chosen as is the case of original DETR (e.g. $N_q = 100$). As we show in the experimental section this contribution alone offers an increment in detection performance over using the content-agnostic object queries of the original implementation. The employment of this module imposes unnoticeable complexities in the overall architecture.

IV. Experimental Study

A. Dataset

In this Section, we present the experimental setup and we discuss the obtained results in the powerline element detection task. To the best of our knowledge there are no publicly available datasets depicting the desired objects or big enough for satisfactory results. Individual researcher efforts so far, such as the CPLID dataset [19] only include a few hundred images and insufficient amount of classes.To this end, we collected and experiment on the AERIAL-CORE powerline inspection dataset. The objects of interest encountered in this dataset are insulators, dumpers and electric towers. Our dataset consists of 11587 images, acquired from video footage of aerial inspection, 8422 of which were used for training and the remaining 3165 for evaluation. Examples of the dataset along with visual results of the experimented detection models, discussed in the following section, are presented in Fig. 2.

B. Experimental setup and results

DETR Resnet50 was chosen as our baseline model. It uses Resnet50 [20] as backbone for image feature extraction and 6 encoder/decoder layers with 8 attention heads each.

The comprehensive results of our experimental study are reported in Table 1. In all of our experiments a COCO-pretrained model was used for capturing general image features, which was then fine-tuned in the AERIA-CORE dataset in order to capture task-specific patterns. A fixed resize on the images was applied at the input layer of the CNN backbone. Specifically, our experiments were conducted using reshaped images of size 448×256. Feeding images of the aforementioned size to the CNN backbone resulted in $\frac{H}{32} \times \frac{W}{32}$ downscale leading to feature maps of size 14×8. As these feature maps will be given as input to the content-specific object query generator based on our previous discussion the output of our query generator will be of temporal sizes $12 \times 6 (= 72)$. We use this result as the number of learnable object queries (i.e. $N_q = 72$) for all of our DETR-based models.

TABLE I
Comparison of state-of-the-art object detectors in the AERIAL-CORE powerline inspection dataset. Average precision (AP) was used as an evaluation metrics.

Model	AP	AP_{50}
SSD [12]	48.3	73.6
Sparse-RCNN [21]	38.9	65.7
DETR [3]	48.7	80.3
DETR† [3]	47.7	75.6
CSOQ-DETR	50.6	81.2
WHIT-DETR	51.1	82.0
CSOQ+WHIT-DETR	**51.6**	**82.5**

At this point, we split our experiments in two stages. In the first stage (top section of Table 1) we compare two state-of-the-art object detectors, namely the Single-Shot multibox Detector (SSD) [12] and the Sparse-RCNN [21] detectors, against baseline DETR models. The same Resnet50 was used

Fig. 2. Image samples of our AERIAL-CORE powerline inspection dataset and elements (e.g. insulators, dumpers and electric towers) detected by DETR detector. The dataset contains a high variety of image shots. Images on the right part of the figure demonstrate the challenges of powerline inspection as some of the objects, e.g. dumpers. are falsely detected or are too small ($32^2 <$ area $< 96^2$ pixels) against an indistinguishable background.

as a CNN backbone in order to establish a more objective comparison between all these different detection methods. SSD and Sparce-RCNN models were trained with all the hyper-parameters values and configurations proposed by the respective authors. For this stage, two baseline DETR models were trained. DETR entry refers to a baseline model with no changes other than those already mentioned. DETR† is of exactly the same architecture except that a instance normalization [22] layer has been applied to the input data before feeding them to the Transformer encoder. All remaining parameters settings (e.g., learning rate, optimizer, dropout etc.) were set equal to the values proposed by the respective authors. By observing the recorded detection metrics we immediately notice that DETR vastly outperforms the its contestants.

In the second stage (bottom section of Table 1) we compare DETR variants to which we have applied our novel contributions. We initially measure the effect of each of our two contributions individually and finally we measure the combined affect. As can be noticed, by solely employing our whitening self-attention variation (WHIT-DETR entry), only at the first layer of the Transformer encoder, we achieve better results than individually incorporating the query generator (CSOQ-DETR entry). However, each of these models manages to outperform the baseline DETR model. Finally, by combining these two contributions we achieve the best recorded performance managing to outperform the baseline DETR model by 2.9% and and 2.2% regarding the AP and AP_{50}, respectively.

V. Conclusions

In this work, we studied a Transformer-based object detector, namely the DETR method and we highlighted certain limitations when specific dataset conditions are met. In order to address those limitations we proposed two novel contributions, the first regarding the attention definition and a second one related with DETR's particular architecture. i.e. a whitened self-attention definition and a content-specific object query generator. Finally, we demonstrated that the incorporation of these two novelties further improves the state-of-the-art in a powerline element detection scenario. Our future work will include studying different definitions of Graph types and how they could be incorporated in different layers in Transformer-based architectures, in more general computer vision problems.

VI. Acknowledgment

This work has received funding from the European Unions Seventh Horizon 2020 research and innovation programme under grant agreement number 871479 (AERIAL-CORE). We would like to thank EDISTRIBUCIN Redes Digitales, S.L. for providing the aerial video footage. This publication reflects the authors views only. The European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] R. Jenssen and D. Roverso, "Intelligent monitoring and inspection of power line components powered by uavs and deep learning," *IEEE Power and energy technology systems journal*, vol. 6, no. 1, pp. 11–21, 2019.

[2] 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 *European conference on computer vision (ECCV)*. Springer, 2014, pp. 740–755.

[3] 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.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems (NIPS)*, 2017, pp. 5998–6008.

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

[7] A. Kessy, A. Lewin, and K. Strimmer, "Optimal whitening and decorrelation," *The American Statistician*, vol. 72, no. 4, pp. 309–314, 2018.

[8] Z.Q. Zhao, P. Zheng, S.t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[9] L. Liu, W. Ouyang, X. Wang, P.l Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[10] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems (NIPS)*, 2016, pp. 379–387.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems (NIPS)*, vol. 28, pp. 91–99, 2015.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision (ECCV)*. Springer, 2016, pp. 21–37.

[13] A. Bochkovskiy, C.Y. Wang, and Hong-Yuan Mark L., "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[14] 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.

[15] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6074–6078.

[16] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146–3154.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning (ICML)*. PMLR, 2015, pp. 448–456.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] X. Tao, D. Zhang, Z. Wang, X. Liu, H. Zhang, and D. Xu, "Detection of power line insulator defects using aerial images analyzed with convolutional neural networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[21] S. Peize, Z. Rufeng, J. Yi, K. Tao, X. Chenfeng, Z. Wei, T. Masayoshi, L. Lei, Y. Zehuan, W. Changhu, and L. Ping, "Sparse r-cnn: End-to-end object detection with learnable proposals," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14454–14463, 2021.

[22] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.