

October 2024

ScaledDETR: An alight weight object detection model for autonomous driving

Ali Elhenidy

A Teaching assistant in computer engineering and control systems, faculty of engineering, Mansoura university, Egypt, alielhenidy@mans.eu.eg

Labib Mohamed

A professeur in computer engineering and control systems, faculty of engineering, Mansoura university, Egypt

Amira Yassien

College Vice Dean for Environmental Affairs and Community Service ,A professeur in computer engineering and control systems, faculty of engineering, Mansoura university, Egypt

Mahmoud saafan

An assosiative professeur in computer engineering and control systems, faculty of engineering, Mansoura university, Egypt

Follow this and additional works at: <https://mej.researchcommons.org/home>



Part of the [Architecture Commons](#), [Engineering Commons](#), and the [Life Sciences Commons](#)

Recommended Citation

Elhenidy, Ali; Mohamed, Labib; Yassien, Amira; and saafan, Mahmoud (2024) "ScaledDETR: An alight weight object detection model for autonomous driving," *Mansoura Engineering Journal*: Vol. 49 : Iss. 5 , Article 13.

Available at: <https://doi.org/10.58491/2735-4202.3238>

This Original Study is brought to you for free and open access by Mansoura Engineering Journal. It has been accepted for inclusion in Mansoura Engineering Journal by an authorized editor of Mansoura Engineering Journal. For more information, please contact mej@mans.edu.eg.

ScaledDETR: An alight weight object detection model for autonomous driving

Ali M. Elhenidy^{*}, Labib M. Labib, Amira Y. Haikal, Mahmoud M. Saafan

Computer Engineering And Control Systems, Faculty of Engineering, Mansoura University, Mansoura ,Egypt

Abstract

End-to-end detection is one of the newest trends in object detection, however it takes a lot of time and memory due to the Transformer encoder-decoder (TED) module. This study proposes ScaledDETR, which implements end-to-end detection based on the most recent efficient backbone with fewer parameters, to address the slow convergence problem in DETR and accelerate the training process. By substituting the effective CNN backbone EfficientNet for the ResNet backbone, ScaledDETR offers an efficient model with fewer parameters. Relative Position Encoding, which has gained 1.3% (AP) improvement, has replaced traditional Position Encoding (PE) in recent times. ScaledDETR employs a single GPU basic architecture that may be useful for applications involving autonomous driving. The suggested model outperforms cutting-edge object detection techniques after only 20 training epochs—a 25-fold reduction from the number of epochs in DETR. In comparison to Faster R-CNN, which achieved 40.2 Ap on the COCO dataset, the suggested technique achieved 41.7 Ap. Using the crowdhuman dataset, ScaledDETR is also assessed and receives 90.12 AP, outperforming PEDETR 89.54 AP, Faster R-CNN 85.0 AP, and Deformable DETR 86.74 AP. When evaluated on the RTX2060 GPU, the model's inference speed is 49 frames per second.

Keywords: Object detection, Deep learning, DETR, Computer vision, CNN

1. Introduction

Object detection is one of the interesting topics in computer vision. Remarkable updates and progress are continuously accomplished to outperform the existing state-of-the-art models. Object detection is a vital task as it has been applied in a wide range of applications autonomous driving (Chen et al., 2020), pedestrian detection, face detection (Topal et al., 2022). Modern object detectors in the deep learning era depend on the recent Convolutional Neural Networks (CNNs) rather than hand-crafted feature designs. Due to the availability of powerful computing devices, object detectors gain a great advance in terms of accuracy and speed. Detection is divided into two tasks; localization of the object and classify it to the corresponding category. Object detection is a challenging task rather than image classification that can't simply use CNN followed by fully connected layers as the number of objects in each image is variant. A simple choice is to extract regions of interest that could contain objects and then feed it to a CNN to identify the object in that region. R-CNN (Girshick et al., 2014) is the first work

to combine CNN with the handcrafted region proposal (Uijlings et al., 2013). Extending this work, (Girshick, 2015, Ren et al., 2016) speeds up the detection pipeline by implementing a Region Proposal Network (RPN) that is a fully convolution module. Dealing with detection tasks in a two-stage manner slows down the pipeline as two branches need to be penalized and fine-tuned to get the final prediction. A series of one-stage with anchor works (Zhao et al., 2019, Bochkovskiy et al., 2020, Tan et al., 2020) propose a novel pipeline that deals with object detection pipelines as a single path.

One of the newest trends in object detection is end-to-end detection; however, because of the Transformer encoder-decoder (TED) module, it requires a lot of time and memory. The first end-to-end object detector with a TED architecture is called Detection Transformer (DETR). End-to-end detectors (Carion et al., 2020, Zhu et al., 2020) remove the post-processing layers such as NMS, and represent objects as a set of sequences using the TED module to get the final prediction in a single path. DETR (Carion et al., 2020) adopts the transformer attention module represented in (Vaswani et al., 2017) used in

Received: 15 April 2024; Revised: 28 June 2024; Accepted: 18 July 2024
Available online 29 October 2024

^{*} Corresponding author. computer engineering and control systems, faculty of engineering, Mansoura university, Egypt, Mansoura, Egypt
E-mail addresses: alielhenidy@mans.edu.eg (A.M. Elhenidy)

<https://doi.org/10.58491/2735-4202.3238>

2735-4202/© 2024 Faculty of Engineering, Mansoura University. This is an open access article under the CC BY 4.0 license (<https://creativecommons.org/licenses/by/4.0/>).

Natural Language Processing (NLP) tasks to the detection pipeline and proves to be effective in computer vision tasks. DETR achieved comparable results with the state-of-the-art object detectors achieving 42.0 AP on COCO test-dev with 41M parameters and 86 Floating Point Operations (FLOPs). Suffering from slow conversion, Deformable DETR (Zhu et al., 2020) adopts a deformable convolution technique to attend to sparse spatial locations by combining the best of the sparse spatial sampling of deformable convolution, and the relation modeling capability of transformers to handle this issue.

Recently, there has been a trade-off between achieving higher accuracy at the expense of time and computation overhead. EfficientDet (Tan et al., 2020) adopts a new family of object detection based on their efficient backbones (Tan and Le, 2019) that achieve comparable results with the state-of-the-art CNN (He et al., 2016, Xie et al., 2017) that adopts a larger scaling architecture with many parameters. Based on one stage detector, EfficientDet (Tan et al., 2020) creates a balanced trade-off between accuracy and efficiency achieving 55.1 AP on the COCO dataset (Lin et al., 2014) with 52M parameters and 226B FLOPs. Efficient models (Howard et al., 2017, Sandler et al., 2018) with a fewer number of parameters and few FLOPs make them good choices for limited resources systems such as mobile and embedded systems applications.

The main contribution of this paper is:

- Propose a novel general object detection model called scaledDETR that uses an efficient and lightweight backbone suitable for autonomous driving applications.
- Relative Position Encoding (Sidonie Carpenter) is adopted rather than standard Position Encoding (PE) to enhance the model performance.
- The number of parameters has been optimized through a parametric study regarding backbone scaling and the number of encoder layers.
- The proposed model achieves 41.7 AP compared with 40.2 AP for the Faster R-CNN baseline and 34.6 AP for the EfficientDet baseline on the COCO dataset.
- ScaledDETR is also assessed and receives 90.12 AP, outperforming PEDETR 89.54 AP, Faster R-CNN 85.0 AP, and Deformable DETR 86.74 AP on the crowdhuman dataset.

The remainder of the paper is arranged as follows; section 2 discusses relevant work. The key components and the suggested architecture are explained in Section 3. The various assessment metrics that are used to assess performance are covered in Section 4. The previous methods and tests will be reviewed in Section 5, along with a comparison of the obtained outcomes with the most

advanced object detectors. Lastly, section 6 introduces the conclusion.

2. Related Work

2.1. End-to-end object detection

Illuminating the hand-designed components is the major objective of modern object detectors. DETR (Carion et al., 2020) was the first work to adopt a transformer that was widely used in machine translation and speech recognition tasks to object detection. A decoder-encoder layer is inserted upon the ResNet backbone to generate a set of predictions that are forced to get a unique map between the predicted bounding boxes and the ground truth via the bipartite matching scheme. Adopting this methodology, the Non-Maximum Suppression (NMS) post-processing is no longer needed and the detection is performed in an end-to-end manner. Figure 1 shows the difference between one-stage models and end-to-end models. Due to the limitation of the transformer attention, DETR requires many training epochs to converge (e.g., 500 epochs). Deformable DETR (Zhu et al., 2020) leverages this drawback by adopting a deformable attention mechanism that attends only to a small set of sampling points. Deformable DETR (Zhu et al., 2020) achieved 43.8AP on the COCO Dataset with only 50 epochs (10X fewer than DETR).

2.2. lightweight object detection models

For various reasons, detecting pedestrians is a difficult problem for original object detectors. Due to overlapping items, most object detectors create duplicate detections for the same object, as the pedestrian frequently demonstrates in crowd scenes. A lightweight object detection model called IterDet (Rukhovich et al., 2021) proposes an iterative approach to detect a fresh subset of objects at each iteration. In order to prevent duplicate detection of the same object, bounding boxes from every iteration are taken into account. Using the NMS to score and eliminate duplicate detections, this method is used with both one- and two-stage detectors. End-to-end detectors (Carion et al., 2020) and (Zhu et al., 2020) have abandoned the manual NMS in favor of treating item detection as a straight-forward set prediction problem. These end-to-end techniques, however, do much worse than traditional two-stage algorithms when it comes to pedestrian identification and crowd scenarios (Lin et al., 2020). In order to make the DETR base architecture acceptable for pedestrian identification, PEDETR (Lin et al., 2020) addresses these shortcomings. In order to precisely control the attention positions, PEDETR suggests a new decoder module called decoder with dense queries and rectified attention field (DQRF), which alters the cross-attention layer. This problem is addressed in another recent work (Zheng et al., 2022),

which suggests a progressive approach to choose queries with high confidence ratings as acceptable queries and then prime them to produce true positive predictions.

2.3. Transformers

Transformers were first adopted in Natural Language Processing (NLP) tasks that deal with sequences efficiently and rely on self-attention. Attention mechanisms mean to give attention to the most important parts of an image and disregard irrelevant parts. (Dosovitskiy et al., 2020) was the first work to replace CNN with a transformer in the scale of

pixels and achieve comparable results on image classification tasks. According to (Guo et al., 2022) Attention mechanisms are categorized according to six categories (e.g., channel attention, spatial attention, temporal attention and branch attention, spatial and temporal attention, and channel and spatial attention). The Transformer (Vaswani et al., 2017) proved to achieve good results by applying attention techniques in machine translation tasks. However, the transformer is a time and memory-consuming technique, many works try to address these limitations.

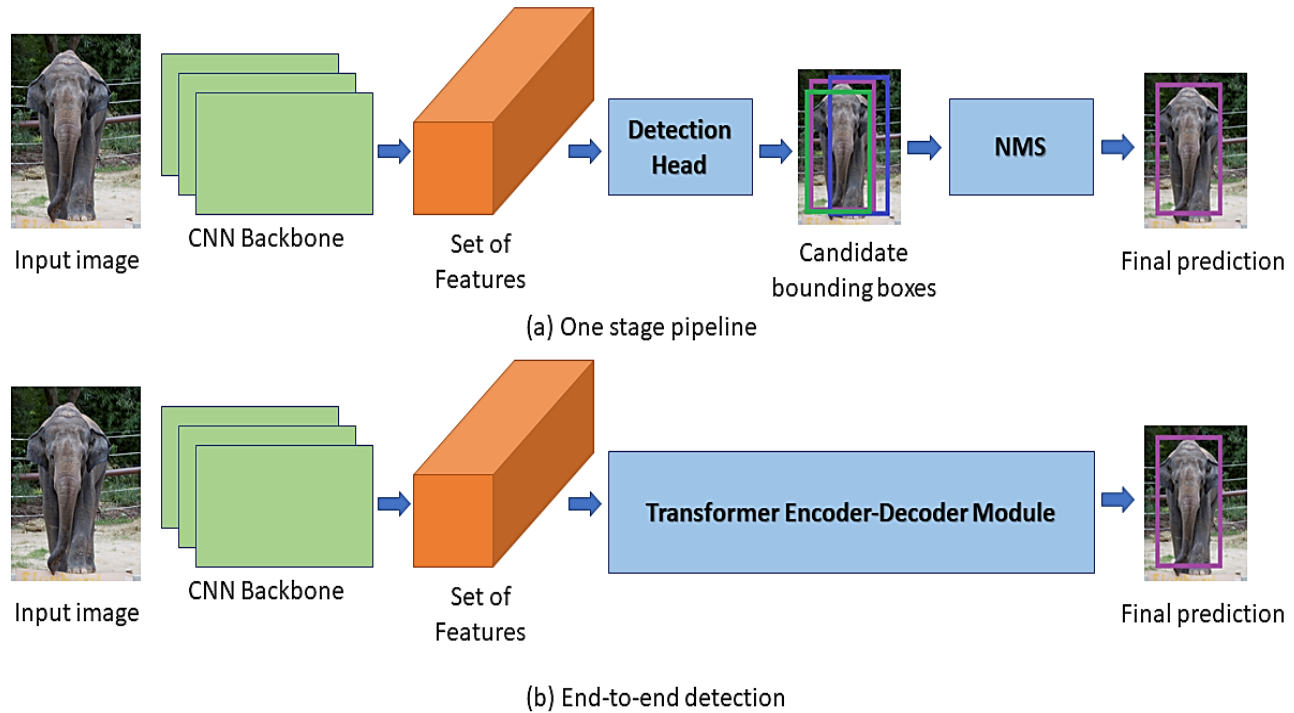


Figure 1. Comparison between one-stage and end-to-end models.

(a) describe the one-stage models that require Non-Maximum Suppression (NMS) to get the final detected bounding box with the highest intersection Over Union (IOU) with the ground truth.

(b) shows the end-to-end pipeline which eliminates the post-processing layer and performs the detection end-to-end across the transformer Encoder-Decoder module that applies bipartite matching between the prediction and the ground truth.

The Vision Transformer (Dosovitskiy et al.) was the first work to apply a pure transformer directly to sequences of image patches images showing that there is no need for CNN anymore. Recent object detector (Touvron et al., 2021) is the first work to insert an attention module in the detection pipeline to perform end-to-end object detection achieving competitive results on the COCO dataset compared with the state-of-the-art baselines, but suffer from low convergence.

2.4. Backbone Scaling

Scaling up the detector baseline is a common technique to obtain better accuracy using bigger backbone models

(e.g., ResNet (He et al., 2016), and ResNeXt (Xie et al., 2017)). There are many recent works to scale up their model whether by increasing the channel size and input image size to 1536x1536 or increasing network width, depth, and resolution (Tan and Le, 2019). A recent work (Touvron et al., 2021), applies attention mechanisms and proposes a simple and more efficient backbone by introducing a split attention module and stacking it with a ResNet block to obtain a new variant. However, ResNeSt-269 (Touvron et al., 2021) has achieved better accuracy than EfficientNet-B7, it drops with 32% less latency.

2.5. Multi-scale Feature Representation

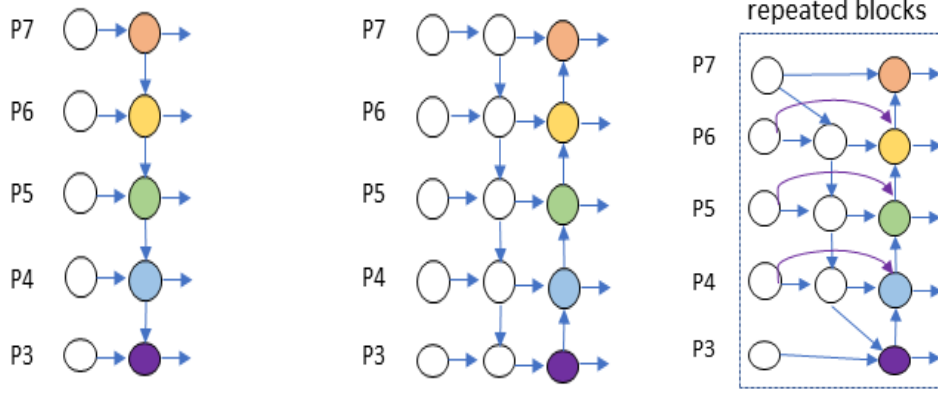


Figure 2. Different feature scaling architectures (a) FPN, (b) PaNet, and (c) BiFPN

One of the challenging tasks in object detection is the scale-invariant problem. Feature Pyramid Network (FPN)(Lin et al., 2017) is one of the works to address this issue by proposing a top-down pathway to combine features from different levels as shown in Figure 2 (a). PANet (Liu et al., 2018) added an extra bottom-up path on top of FPN as shown in Figure 2 (b). EfficientDet (Tan et al., 2020) tried to optimize multi-scale feature fusion by proposing a BiFPN block as shown in Figure 2 (c).

Feature fusion has been popular and essential in many detectors and backbone architectures to obtain more representative information and gain performance enhancement. M2det (Zhao et al., 2019) proposes a U-shape module that fuses extracted features by the backbone from multi-levels. A feature pyramid is developed upon the decoder layers for detecting objects.

3. ScaledDETR Model

ScaledDETR is an end-to-end object detector that relies on a minimally parameterized, effective backbone. The various model components will be explained in this

section. In order to extract a set of visual features, it begins with the first image, which is sent to CNN. The features appended to the spatial positional encodings are flattened by the model. After receiving the positional encodings as input, the encoder layer used global scene reasoning to separate objects. The transformer decoder receives the encoder layer's output and applies a fixed number of learnable object queries (N) (N = 100) to it. Three feed-forward networks (FFNs) compute the final forecast. Figure 3 shows the entire architecture.

3.1. Backbone

EfficientNet (Tan and Le, 2019) is used as a backbone which consists of 7 mobile inverted bottleneck (MBConv) blocks. The last fully connected layer is removed and flattened to construct positional encodings upon it. EfficientNet requires 7.8M parameters which is time and memory-efficient compared with ResNet-50 used in DETR which requires 26M parameters. Backbone scaling has a major effect on the number of the model parameters the training and inference time and the number of the floating point operations consequently.

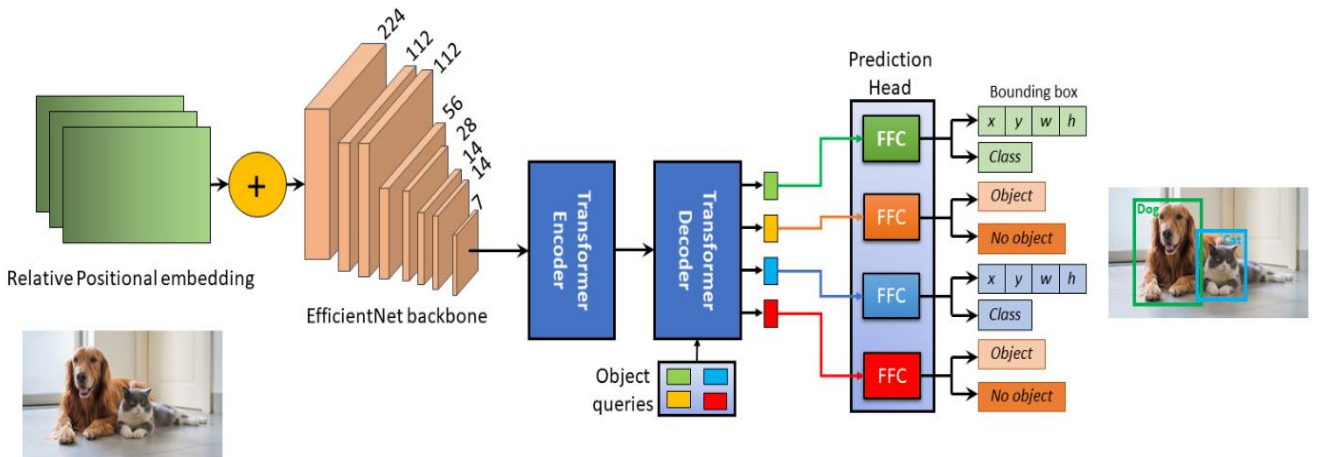


Figure 3. ScaledDETR architecture

3.2. Transformer attention

The same Transformer encoder-decoder (TED) module presented by (Carion et al., 2020) is used in the presented work. The main core of the TED module is the head self-attention layer. The attention mechanism maps a query to a certain value using a set of key pairs as shown in Equation 1. The head self-attention layer is repeated h times in a parallel manner and concatenated to output the final

attention. In the proposed model, six encoder-decoder layers are built upon the image features to get the final class and bounding box prediction. The Transformer encoder-decoder architecture is shown in Figure 4.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (1)$$

Where Q is the query, K is the value, V is the value, and k is the dimension of the key.

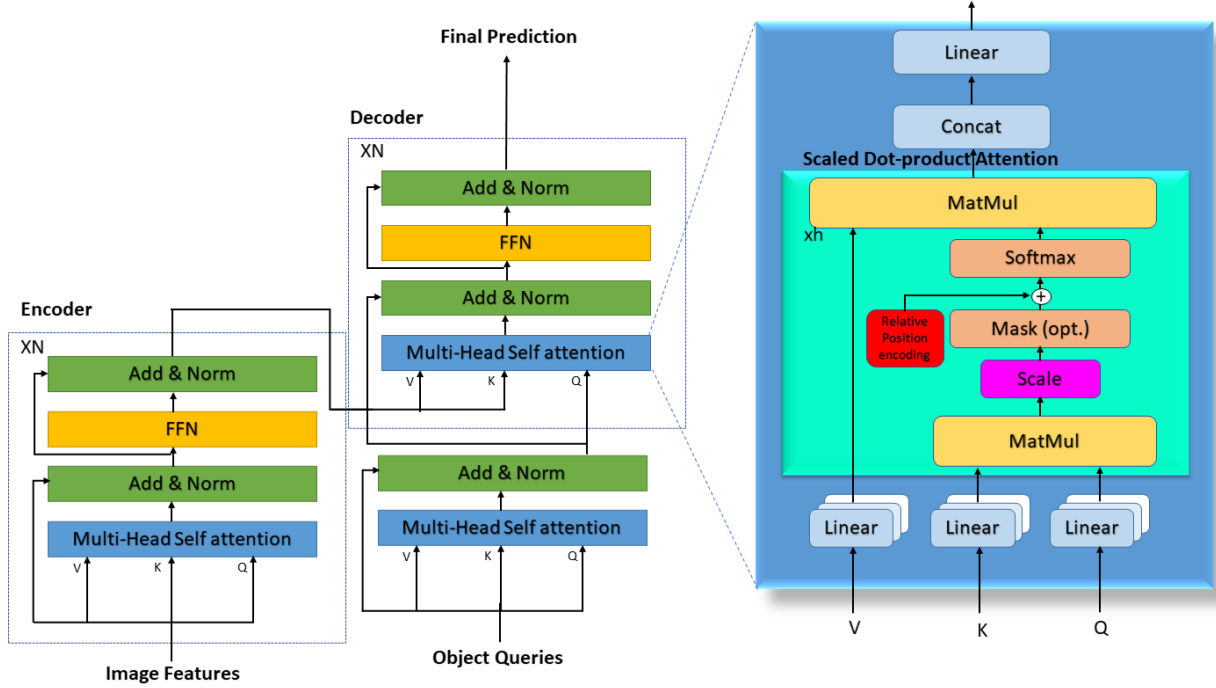


Figure 4: Transformer encoder-decoder (TED) architecture

3.3. Position Encoding

Position Encoding is a learnable token that forces the model to consider the order of the input tokens. DETR adopts Absolute Position Encoding (APE) and simply adds the image features to the position encoding to make use of the order of the sequence. Image features are unrolled to three equal components; value, key, and query. Position Encoding is added to the key and the query at the encoder stage and added to the key in the decoder stage. The three components are packed into the 1D vector to represent a sequence. A matrix multiplication between the key and the query is computed and scaled over the dimension of the key. In DETR APE is computed by fixed encodings using sine and cosine functions with different frequencies. Relative position encoding (Wu et al., 2021) takes into account both the interactions between queries and relative position embeddings as well as directional relative distance modeling. The suggested RPE techniques are lightweight and easy to use. Plugging them into transformer blocks is a simple process.

3.4. Loss function

Since the end-to-end object detector infers predictions in a single pass, DETR first uses a bipartite matching between the prediction and the ground truth. The loss function is computed using the Hungarian loss (Kuhn, 1955) for all pairs matched that gets the final detection of the class and the bounding box in a single assignment as shown in Eq (2)

$$L_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^N [-\log p^{\hat{\sigma}(i)}(C_i) + 1_{\{C_i \neq \emptyset\}} L_{\text{box}}(b_i, \hat{b}_{\sigma^{\wedge}}(i))] \quad (2)$$

where σ^{\wedge} is the optimal assignment of the detection compared with the ground truth, $p^{\hat{\sigma}(i)}(C_i)$ is the probability of class C_i , $\hat{b}_{\sigma^{\wedge}}(i)$ is the predicted box, and L_{box} is the bounding box loss which is computed using a linear combination of the L1 loss and the generalized IoU loss (Rezatofighi et al., 2019) which is computed according to Eq 3

$$L_{\text{box}} = \lambda_{\text{iou}} L_{\text{iou}}(b_i, \hat{b}_{\sigma(i)}) + \lambda L1 \|b_i - \hat{b}_{\sigma(i)}\| \quad (3)$$

Where λ_{iou} and $\lambda L1$ are hyperparameters.

An auxiliary loss is proved effective in the decoder during training, especially to help the model output the correct number of objects of each class.

4. Performance Metrics and dataset

This section will cover the utilized performance metrics and the data benchmark.

4.1. Metrics

Object detection algorithms not only classify an object, but also, they localizing it. The output of the model is divided into two branches; the classification layer and the bounding box regressor to localize the corresponding object. The performance of the model is evaluated corresponding to how confident the model prediction is. The confidence score represents how the predicted box is overlapped with the ground truth. This is calculated by the Intersection Over Union (IOU) metric as the area of intersection divided by the area of their union as shown in Figure 5. Recall and precision are two significant variables that are evident, according to IOU. Recall is the proportion of true positive instances to all relevant cases (all ground-truth bounding boxes), whereas precision is the percentage of true positive examples to all positive predictions.

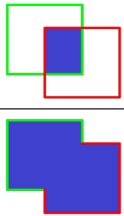
$$\text{IOU} = \frac{\text{area of overlap}}{\text{area of union}} = \frac{\text{Figure 5}}{\text{Figure 5}}$$


Figure 5. Intersection Over Union (IOU) metric

The COCO dataset includes metrics such as Average Precision (AP) and Average Recall (AR) that are pertinent to object detectors. Recall and precision are used to create the AP metric, which ranks retrieval results. The six values that constitute AP are AP@ [0.5:0.95] together with APS, APM, and APL. The AP with ten distinct IOU thresholds ($t = [0.5, 0.55, \dots, 0.95]$) is calculated to obtain the AP@.5 and AP@.75 measures. The average of all computed values is then taken. Ground-truth objects are evaluated by APS, APM, and APL based on the area size (that is, area < 322 pixels, area < 962 pixels, and area > 962), respectively. The average AP for every class is called Mean Average Precision. Finding the Average Precision (AP) for each class and averaging it over several classes yields the mAP as shown in Eq 4. The trade-off between recall and precision is taken into account by the mAP, which also takes false positives (FP) and false negatives (FN) into account. Because of this feature, mAP is a useful statistic for the majority of detection applications.

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N \text{AP}_i \quad (4)$$

Log-average miss rate (MR-2) metric calculates the miss rate on false positives for each image on a log-scale, with a range of [10⁻², 100]. This metric is used to evaluated the performance on the crowdhuman dataset (Shao et al., 2018), because it illustrates the number of pedestrians that are missed.

4.2. Dataset

The proposed method is trained and evaluated on the COCO (Lin et al., 2014) dataset stands for Common Objects in Context. The COCO dataset was published by Microsoft in 2014 and is applied in many computer vision tasks such as object detection, captioning, and segmentation. There are two versions of the COCO dataset, the first released version in 2014 and the updated version in 2017. The labeled image classes in the two versions in the same which is 91 labeled classes, but the annotated classes for object detection are 80 classes. The difference between the two versions is the amount of data for the train and the test. COCO 2014 version contains 83k images for train and 41k for validation and test, whereas COCO 2017 contains 118k images for train, 5k for validation, and 41k for test. ScaledDETR is trained and tested on the COCO 2017 version.

The CrowdHuman dataset (Shao et al., 2018) is much more challenging as it contains more instances per images, and those instances are often highly overlapped. The dataset contains 470K human instances with various types of occlusions, with 22.6 people per image from the train and validation subsets. Human head bounding-box, human visible-region bounding-box, and human full-body bounding-box annotations are applied to each person instance.

5. Simulation analysis and results

All experiments are trained and tested using a single RTX 2060 GPU with 6G NVRAM. The corresponding environment is Python 3 and pytorch 1.10.1 with Cuda 10.2 version that contains the corresponding optimizer. To import the EfficientNet backbone, the Timm library is installed. EfficientDETR Follow the parameters setting for DETR (Carion et al., 2020) transformer's learning rate is set to 10⁻⁴, weight decay to 10e-4, and the backbone's learning rate to 10e-5. The 6 encoder-decoder layers are learned through 100 object queries (N=100), where N is larger than the number of objects.

The ScaledDETR model is trained on the COCO 2017 detection dataset for only 20 epochs.

ScaledDETR is evaluated on the COCO dataset with 118K images for training and 5K for testing. The proposed model achieves 40.4 AP -as shown in Table 1 – compared with 40.2 AP Faster R-CNN (Ren et al., 2016) baseline and 42.0 AP for DETR (Carion et al., 2020). The proposed model achieves 41.7 Ap with (+6.5%) higher than the EfficientDet baseline with only 25.6M parameters.

Results show that ScaledDETR is (1.6%) behind the DTER model, but the proposed architecture is more efficient with 23.8M parameters compared with 41.3M to the latter. The model is trained using Stochastic Gradient Descent (SGD) rather than AdamW (Loshchilov and Hutter, 2017) used in DETR and the model for only 20

epochs which are 25x fewer than the number of epochs in DETR and achieves competitive results with the state-of-the-art object detection methods. The recent Relative Position Encoding is adopted rather than standard Position Encoding which proved to gain a 1.3% (AP) improvement.

Table 1. Comparison with Faster R-CNN, EfficientDet, and DETR on COCO 2007 test-dev

Method	#Parameters	Backbone	AP	AP _{s0}	AP ₇₅	AP _s	AP _M	AP _L
Faster R-CNN (Ren et al., 2016)	56M	VGG-16	21.9	42.7	-	-	-	-
EfficientDet-D0 (512) (Tan et al., 2020)	32M	EfficientNet	34.6	53.0	37.1	-	-	-
EfficientDet-D1(640) (Tan et al., 2020)	32M	EfficientNet	40.5	59.1	43.7	-	-	-
DETR (Carion et al., 2020)	41.3M	Residual-101	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5 (Carion et al., 2020)	41.3M	Residual-101	43.3	63.1	45.9	22.5	47.3	61.1
DETR-DC5-R101 (Carion et al., 2020)	41.3M	Residual-101	44.9	64.7	47.7	23.7	49.5	62.3
YOLOv4 (Bochkovskiy et al., 2020)		Darknet53	44.6	64.1	49.5	27	49	65.7
The proposed ScaledDETR	23.8M	EfficientNet	40.4	58.3	44.1	19.4	45.0	61.0
The proposed ScaledDETR (with RPE)	25.6M	EfficientNet	41.7	60.1	44.6	20.2	45.6	61.3

Table 2. Comparison with Faster R-CNN, Deformable DETR, and PEDETR on the crowdhuman dataset

Method	AP	MR ²
Faster R-CNN (Ren et al., 2016)	85.0	50.4
Deformable DETR (Zhu et al., 2020)	86.74	53.98
PEDETR (Lin et al., 2020)	89.54	45.57
The proposed ScaledDETR	90.12	48.2

Table 3. Comparison of inference speed with lightweight object detection models (YOLOv4 and YOLOv7)

Method	#Parameters	FPS
YOLOv4 (Bochkovskiy et al., 2020)	64.4M	41
YOLOv7 (Wang et al., 2023)	36.9M	161
The proposed ScaledDETR	25.6M	49

Table 4. Comparison of inference speed on different architectures

Architecture	FPS
RTX2060 GPU	49
intel core i7 @ 2.6 GHZ	25
Jetson nano GPU @ 128 cores	41

ScaledDETR can generalize to different benchmarks and achieve competitive results. ScaledDETR is also evaluated on the crowdhuman dataset as listed in Table 2 and achieves 90.12 AP compared with Faster R-CNN 85.0 AP, Deformable DETR 86.74 AP, and PEDETR 89.54 AP. The miss rate

MR² for the proposed model is less than Faster R-CNN and, Deformable DETR with a gain of 1.6 and 6.2 respectively. ScaledDETR is tested against other lightweight object detection models to evaluate the inference speed for real time applications as listed in Table 3. ScaledDETR achieves 49 frame per second (FPS) against 41 FPS for YOLOv4 and 161 FPS for YOLOv7. (Wang et al., 2023) However, YOLOv7 has the heights number of parameters compared with YOLOv4 and ScaledDETR, it has the heights inference speed.

Table 5. Comparison between Dter models based on EfficientNet and Resnet backbones variants from efficiency, accuracy, and Floating point computations points of view on the COCO dataset

Model	No. Parameters	FLOPs	mAP
ResNet18	28.68M	43.7B	35.5
ResNet34	38.71M	51.9B	37.1
ResNet50	41.31M	64.1B	38.6
ResNet101	60.24M	72.5B	38.9
ResNet152	75.83M	140.2B	39.2
efficientnet_b0	21.22M	8.3B	35.4
efficientnet_b1	23.73M	9.1B	35.9
efficientnet_b2	24.83M	10.5B	36.2
efficientnet_b3	27.73M	12.6B	36.8
efficientnet_b4	34.37M	16.8B	38.4
efficientnet_b5	44.91M	17.6B	38.9
efficientnet_b6	57.03M	19B	39.5
efficientnet_b7	79.77M	23.9B	40.2

The proposed model is tested on a wide range of hardware configurations. The model inference speed is tested on the RTX2060 GPU, the intel core i7 processor with 2.5 GHZ and the jeston nano kit with the NVIDIA Maxwell architecture 128 core. The model is tested on both static images and video streams. ScaledDETR has 49 frame per second (FPS) inference speed on the RTX 2060 GPU, 25 FPS on intel core i7@ 2.6 GHZ processor, and 41 FPS on Jetson nano GPU @128 cores. The reported results are listed in Table 4.

A further discussion on the effect of the efficient backbone is listed in Table 5. The efficient variants are compared with the ResNet variants from the efficiency, accuracy, and Floating point computations points of view. It is noticed that efficientnet backbone is more efficient in terms of number of parametrs on the floating point operations.

6. Conclusions

In this paper, ScaledDETR is presented as an end-to-end object detection method that makes use of the efficientNet CNN backbone with fewer parameters. ScaledDETR handles the problem of memory and time consumption of the end-to-end methods due to the encoder-decoder transformer module that has to keep attention to a long sequence with large numbers of parameters. With ScaledDETR architecture, the model is trained and tested on a single GPU with fewer numbers of epochs. The model is evaluated on the COCO dataset and achieve competitive results compared with Efficientdet and DETR baseline. The proposed model achieves 41.7 Ap with (+6.5%) higher than the EfficientDet baseline with only 25.6M parameters.

ScaledDETR can generalize to different benchmarks and achieve competitive results. ScaledDETR achieves 90.12 AP compared with Faster R-CNN 85.0 AP, Deformable DETR 86.74 AP, and PEDETR 89.54 AP on the crowdhuman dataset. The proposed methodology proves that an efficient backbone is an essential key in optimizing the detection algorithm in terms of the number of parameters and floating-point operations.

Author contributions statement

The corresponding author (Eng. Ali M. Elhenidy) is responsible for coding the proposed model and writing the paper.

The second author (Prof. Labib M. Labib) is responsible for providing the data, and discussing the results.

The third author (Prof. Amira Y. Haikal) is responsible revising and helping with the writing of the manuscript.

The fourth author (Assoc. Prof. Mahmoud M. Saafan) is responsible for plagiarism check and paraphrasing and helping the first author in coding the software.

Conflicts of interest

There are no conflicts of interest.

References

- BOCHKOVSKIY, A., WANG, C.-Y. & LIAO, H.-Y. M. J. A. P. A. 2020. Yolov4: Optimal speed and accuracy of object detection.
- CARION, N., MASSA, F., SYNNAEVE, G., USUNIER, N., KIRILLOV, A. & ZAGORUYKO, S. End-to-end object detection with transformers. European conference on computer vision, 2020. Springer, 213-229.
- CHEN, G., CAO, H., CONRADT, J., TANG, H., ROHRBEIN, F. & KNOLL, A. J. I. S. P. M. 2020. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. 37, 34-49.
- DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEGHANI, M., MINDERER, M., HEIGOLD, G. & GELLY, S. J. A. P. A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale.
- GIRSHICK, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015. 1440-1448.
- GIRSHICK, R., DONAHUE, J., DARRELL, T. & MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2014. 580-587.
- GUO, M.-H., XU, T.-X., LIU, J.-J., LIU, Z.-N., JIANG, P.-T., MU, T.-J., ZHANG, S.-H., MARTIN, R. R., CHENG, M.-M. & HU, S.-M. J. C. V. M. 2022. Attention mechanisms in computer vision: A survey. 8, 331-368.
- HE, K., ZHANG, X., REN, S. & SUN, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 770-778.
- HOWARD, A. G., ZHU, M., CHEN, B., KALENICHENKO, D., WANG, W., WEYAND, T., ANDREETTO, M. & ADAM, H. J. A. P. A. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications.
- KUHN, H. W. J. N. R. L. Q. 1955. The Hungarian method for the assignment problem. 2, 83-97.
- LIN, M., LI, C., BU, X., SUN, M., LIN, C., YAN, J., OUYANG, W. & DENG, Z. J. A. P. A. 2020. Dettr for crowd pedestrian detection.
- LIN, T.-Y., DOLLÁR, P., GIRSHICK, R., HE, K., HARIHARAN, B. & BELONGIE, S. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 2117-2125.
- LIN, T.-Y., MAIRE, M., BELONGIE, S., HAYS, J., PERONA, P., RAMANAN, D., DOLLÁR, P. & ZITNICK, C. L. Microsoft coco: Common objects in context. Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, 2014. Springer, 740-755.
- LIU, S., QI, L., QIN, H., SHI, J. & JIA, J. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 8759-8768.
- LOSHCHILOV, I. & HUTTER, F. J. A. P. A. 2017. Decoupled weight decay regularization.
- REN, S., HE, K., GIRSHICK, R., SUN, J. J. I. T. O. P. A. & INTELLIGENCE, M. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. 39, 1137-1149.
- REZATOFIGHI, H., TSOL, N., GWAK, J., SADEGHIAN, A., REID, I. & SAVARESE, S. Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019. 658-666.
- RUKHOVICH, D., SOFIIUK, K., GALEEV, D., BARINOVA, O. & KONUSHIN, A. Iterdet: iterative scheme for object detection in crowded environments. Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2020, Padua, Italy, January 21-22, 2021, Proceedings, 2021. Springer, 344-354.
- SANDLER, M., HOWARD, A., ZHU, M., ZHMOGINOV, A. & CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 4510-4520.
- SHAO, S., ZHAO, Z., LI, B., XIAO, T., YU, G., ZHANG, X. & SUN, J. J. A. P. A. 2018. Crowdhuman: A benchmark for detecting human in a crowd.
- SIDONIE CARPENTER 2014. *Growing Green Guide: A guide to green roofs, walls and facades*.
- TAN, M. & LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning, 2019. PMLR, 6105-6114.
- TAN, M., PANG, R. & LE, Q. V. Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020. 10781-10790.
- TOPAL, B. B., YURET, D. & SEZGIN, T. M. J. A. P. A. 2022. Domain-adaptive self-supervised pre-training for face & body detection in drawings.
- TOUVRON, H., CORD, M., DOUZE, M., MASSA, F., SABLAYROLLES, A. & JÉGOU, H. Training data-efficient image transformers & distillation through attention. International conference on machine learning, 2021. PMLR, 10347-10357.
- UIJLINGS, J. R., VAN DE SANDE, K. E., GEVERS, T. & SMEULDERS, A. W. J. I. J. O. C. V. 2013. Selective search for object recognition. 104, 154-171.
- VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. & POLOSUKHIN, I. J. A. I. N. I. P. S. 2017. Attention is all you need. 30.
- WANG, C.-Y., BOCHKOVSKIY, A. & LIAO, H.-Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time

- object detectors. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023. 7464-7475.
- WU, K., PENG, H., CHEN, M., FU, J. & CHAO, H. Rethinking and improving relative position encoding for vision transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10033-10041.
- XIE, S., GIRSHICK, R., DOLLÁR, P., TU, Z. & HE, K. Aggregated residual transformations for deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 1492-1500.
- ZHAO, Q., SHENG, T., WANG, Y., TANG, Z., CHEN, Y., CAI, L. & LING, H. M2det: A single-shot object detector based on multi-level feature pyramid network. Proceedings of the AAAI conference on artificial intelligence, 2019. 9259-9266.
- ZHENG, A., ZHANG, Y., ZHANG, X., QI, X. & SUN, J. Progressive end-to-end object detection in crowded scenes. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022. 857-866.
- ZHU, X., SU, W., LU, L., LI, B., WANG, X. & DAI, J. J. A. P. A. 2020. Deformable detr: Deformable transformers for end-to-end object detection.