Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

Safety Helmet Detection on Field Project Worker Using Detection Transformer

Muhammad Rayhan Subhi, Ema Rachmawati*, Gamma Kosala

School of Computing, Telkom University, Bandung
Jl. Telekomunikasi. 1, Terusan Buahbatu - Bojongsoang, Telkom University, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat, Indonesia
Email: [1]mrayhans@telkomuniversity.ac.id, [2,*]emarachmawati@telkomuniversity.ac.id,
[3]gammakosala@telkomuniversity.ac.id
Correspondence Author Email: emarachmawati@telkomuniversity.ac.id
Submitted: **13/07/2023;** Accepted: **28/07/2023;** Published: **31/07/2023**

Abstract−There have been many cases of work accidents caused by not complying with safety standards at work, especially in the use of safety helmets. This study is able to make regular observations in identifying project personnel using safety helmets at work, this aims to reduce the risk of accidents at work, namely in the use of helmet attributes at work. Some previous studies, have proposed the use of image detection-based models using the Detection Transformer (DeTr) method for obtaining object detection, group prediction, and combining methods, using the Intersection over Union (IoU) method for obtaining object detection results, to achieve the best performance, namely to get convergence results. Based on the combination of these two methods, the results value of average IoU is 0.50 from 500 identified project personnel data were obtained.

Keywords: Safety Helmet Detection; IoU; DeTr; Deep Learning

1. INTRODUCTION

Along with the awareness of the importance of security in substations monitoring systems have become very crucial. In recent decades artificial intelligence technologies such as computer vision and machine learning have been widely applied in the development of substation intelligent monitoring. The continuous progress in this field has brought significant benefits in improving the monitoring capability and reliability of substation systems [1]. Construction work or field project work is one of the jobs that has a high risk of accidents. In project work There are many risks of accidents that can occur, such as being hit by falling objects, bumped, slipped, tripped, hit by sharp objects, and others. Also, the head is the most important part of the body that must be protected from various kinds of accident factors [2],.

In addition, one of the accident factors found in field projects is the lack of awareness of workers to use safety equipment. This can actually be overcome by using personal protective equipment. One of the personal protective equipment is a safety helmet that can reduce the rate of accidents at work. Safety helmet is one of the personal protective equipment that has a function to protect the head from various objects so that the head is not injured. However, the negligence rate of project workers in using helmets is still high, causing a very high risk of accidents. To overcome the problems that occur, a system is created to detect the use of safety helmets in field project workers.

Research conducted by Wang et al.(2020)[3], is to detect safety helmets using the CSYOLOv3 method. In their study, experiments were conducted under different conditions, including crowds and small targets. The results of their research show the accuracy rate of safety helmet users is high, which is an average of 90%. Another study conducted by Hayat et al.(2022)[4] , they created a system using the YOLO method which has high speed and can process 45 frames per second, their work getting an accuracy of 92.44% in detecting objects in smaller with low light object. Research that has been conducted by Lin et al.(2021) regarding crowd detection using the Detection Transformer (DeTr) method[5], the results of work have a high level of accuracy in object detection. In their study, they developed a pedestrian crowd detection system using the CityPersons dataset, consisting of 2,975 images for training and 500 images for validation, as well as the CrowdHuman dataset, consisting of 15,000 images for training and 4,370 images for validation. The results of their research showed high accuracy in pedestrian detection using the DeTr method. The three studies mentioned, it can be seen that their research wants to detect an object in the form of a safety helmet that has a function for protection in the field project work environment.

Another studies were done by Rescky et al.(2022) [6]. Their study used YOLO and CNN methods for detection safety vests and helmets. The results of their work demonstrate a good measure of detection speed and accuracy. On the other hand, the modified CNN method shows an average accuracy of 90%. A drawback of their system is that it cannot detect all head and body objects in the image during the test. Investigation on helmet detection was performed by Setyawan et al.(2021) [7]. In their study, the authors used YOLO V3 method to create a no helmets for motorcycle and excess passenger detection system. The dataset used in their study was 173 images consisting of motorcycles, helmets, no helmets, riders, or people. The results of their study shows a good accuracy rate of 84.6%. A drawback of their work is that errors are still present in the images when riders without helmets and wearing accessories such as hats and helmets.

The purpose of this study is to design a system that can detect whether the field project worker is wearing a safety helmet or not. In this study, an object detection system is built using the DeTr that is one of the methods

Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

of deep learning methods used to detect an object for getting higher accuracy result by compared to previous research. The results of this study are expected to increase awareness on field project workers to use safety helmets in carrying out tasks, and also expected to reduce the level of accidents or injuries to project workers who are carrying out their duties in the field or injury to project workers who are carrying out their duties in the field.

Based on these problems, we aim to conduct a study on the use of the Detection Transformer method with the safety helmet dataset to detect the use of safety helmets in field project workers. We hope that this study can be a model for developing an effective and accurate system to be able to detect the use of safety helmets among field project woerkers using the Detection Transformer method.

2. RESEARCH METHODOLOGY

2.1 System Design

The system implemented in this study consists of several phases, as shown in Figure 1.

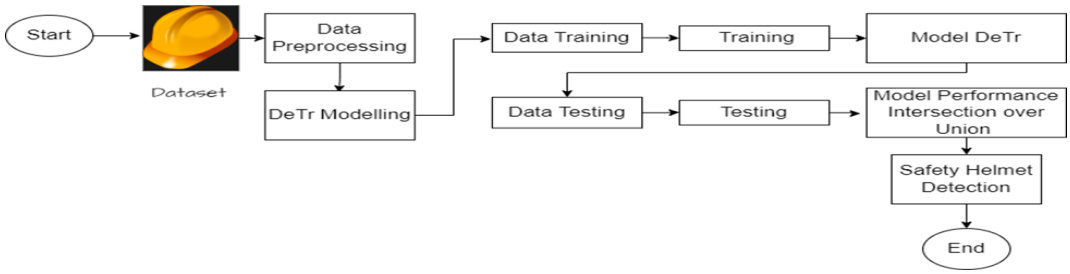

Figure 1. System Design

The study will involve several phases, beginning with the collection and preparation of a safety helmet dataset. Data preprocessing techniques will be employed to ensure that the dataset is suitable for training the DeTr model. Subsequently, the DeTr model will be trained using the prepared dataset, and testing will be conducted to evaluate its performance in detecting safety helmets. The performance of the model will be measured using the Intersection over Union (IoU) metric, which assesses the accuracy and effectiveness of the safety helmet detection system.

2.2 Preprocessing

Preprocessing is a phase performed on an image to change the value of an image [8]. At this phase, the researcher only changed the image size to 224x224. This resizing is done to speed up the training process by reducing computational complexity. By resizing the images the next phase of training can be performed more efficiently.

2.3 DeTr Modelling

The Detection Transformer (DeTr) model was introduced in the paper "End-to-End Object Detection with Transformers" by Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. This approach enables end-to-end training for object detection, simplifying many of the complexities found in models such as Faster R-CNN and Mask R-CNN. These complexities include area proposals, non-maximum suppression procedures, and anchor generation. Additionally, DeTr can easily be extended to perform panoptic segmentation in a unified manner. Then, DeTr is a recently proposed Transformer-based method that considers object detection as an ensemble prediction problem and achieves state-of-the-art performance but requires particularly long training times for convergence[9]. DeTr is a detector that can perform object detection using CNNs using the "Transformer" model and can remove some components such as non-maximal suppression and anchors [10]. The model is implemented directly on the prediction object and ground truth using bipartite matching, and the truth is implemented using bipartite matching and the Hungarian algorithm [11]. The architecture of DeTr is shown in Figure 2 [11].

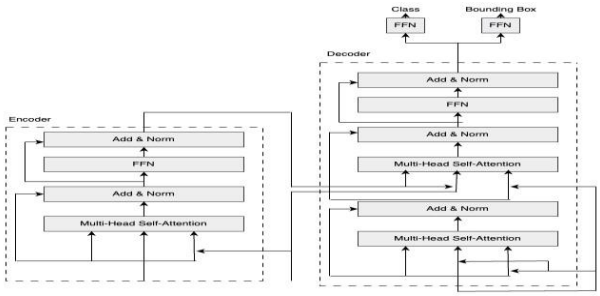

Figure 2. Architecture of DeTr[11]

Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

Compared to other object detection models such as Faster R-CNN or Mask R-CNN, DeTr adopts a simpler approach by using few hyperparameters. In DeTr, there is no need to set parameters such as the number of boxes (region proposal), aspect ratio, bounding box coordinates or default value coordinates, as well as the non-maximum suppression (NMS) boundary. This approach eliminates those complex steps and directly applies the encoder-decoder transformer in the model to achieve a more general and versatile approach for various applications. In other words, DeTr simplifies the object detection process by using a transformer model, reduces the need for complex parameter settings, and enables wider adaptation to various object detection tasks [12]. DeTr, or Detection Transformer, uses a sequence of observations to perform prediction in parallel. The DeTr architecture consists of a Convolutional Neural Network (CNN) as the main part that extracts visual features from the input image. These features are then processed by the Transformer model which consists of an encoder and decoder. During training, DeTr performs a two-stage training process. The first stage is to predict the class label and bounding box coordinates for each object in the image. At this stage, the model is paired with ground truth boxes using a Hungarian algorithm that selects the best box pairs based on their similarity. The second stage is to handle the unpaired predictions. This happens when there are more box predictions than the ground truth box or when the IoU (Intersection over Union) between the box predictions and the ground truth box is below a certain threshold. In such cases, the model labels a special class "no object" to the unmatched predictions. This helps the model to distinguish between true objects and background or false detections. By performing this two-stage training, DeTr can cope with situations where the number of objects in the image may vary, and produce clear predictions by explicitly labelling the unpaired predictions with "no object". Figure 3 shows the pipeline of DeTr [11]

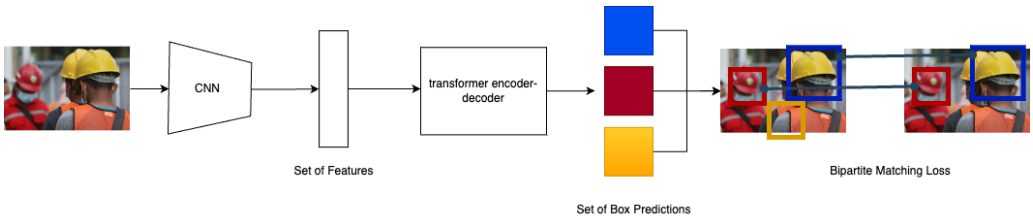

Figure 3. The Pipeline of DeTr[11]

During training, the dataset undergoes a modelling process where images are classified based on available training and validation data. To facilitate training, the original images in the dataset are divided into different patch sizes. This allows the model to study and analyze different regions of the image and increase the understanding of different objects and their properties. The system modeling process is shown in Figure 4.

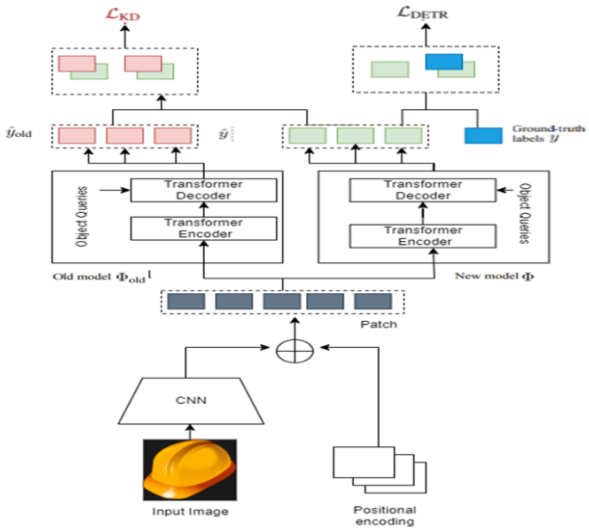

Figure 4. DeTr Modelling Process[13]

The Encoder-Decoder Transformer architecture is a powerful model commonly used in a variety of tasks, including image processing. When an image is fed into an Encoder-Decoder, it passes through a series of layers that make up the Encoder [14]. The Transformer Encoder-Decoder architecture consists of multiple self-attention layers and a feed-forward neural network in the Encoder. These layers transform image features by attending to different parts of the image, capturing spatial relationships and dependencies. The output of each self-attention layer is then passed through the feed-forward network for further processing. In the Decoder, which also has self-attention and feed-forward layers, a learned position embedding called an object query is introduced as an additional input. These object queries represent representations of specific learned objects or regions in the image. They allow the Decoder to focus on the relevant part of the image when generating the output. By attending to

Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

different object queries, the Decoder selectively processes information related to a specific object or region. The output of the Encoder and object queries are combined and passed through the Decoder layer. The Decoder learns to pay attention to relevant image features and object queries to generate the desired output, such as image classification, object detection, or image generation[11]. The encoder and decoder is shown in Figure 5.

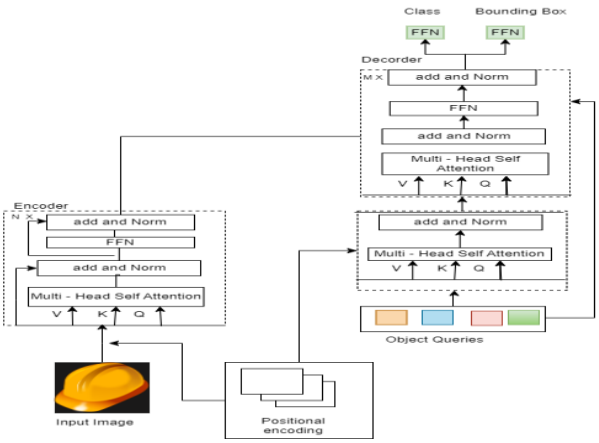

Figure 5. Encoder and Decorder[13]

DeTr uses a ResNet-50 or ResNet-101 CNN backbone trained with ImageNet. The DeTr and DeTr-R101 models have an output with the number of channels C=2048 and size H,W=H0/32, W0/32. Dilation can be applied at the last stage of the backbone to increase the feature resolution. The corresponding models are called DeTr-DC5 and DeTr-DC5-R101 (dilated C5 stage). After levelling the representation and equipping it with positional encoding, the model feeds into the Transformer encoder. The model has 6 encoder and 6 decoder layers with 256 width and 8 heads of concern. Each decoder output embedding is passed to a shared feed forward network (FFN) that predicts the detection or "no object" class. Additional loss is added to the decoder during training to help the model output the correct number of objects from each class. All prediction FFNs share their parameters and a shared Norm Layer is used to normalise the inputs to the prediction FFNs from different decoder layers[15].

Unlike other object detection methods that rely on matching multiple predicted bounding boxes with a single ground truth box. DeTr uses a two-part or one-to-one matching strategy. This approach is different from the traditional bipartite matching used in other methods. By using one-to-one matching, DeTr can effectively reduce the number of poor quality predictions and reduce the performance penalty associated with techniques such as non-maximum suppression (NMS). The two-part matching loss is calculated using the Hungarian algorithm, and the overall DeTr loss is calculated based on the two-part matching loss. The specific formula for calculating DeTr loss can be derived from the equations used in this context. This novel approach in matching and loss calculation contributes to significant performance improvements and eliminates the drawbacks usually associated with the NMS technique (1).

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_i^N \quad \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \tag{1}$$

where $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is a pair-wise matching cost between ground truth yi and a prediction with index σ(i). The matching cost takes into account various factors such as the predicted class labels and the similarity between the predicted bounding boxes and the ground truth boxes. The goal is to find the assignment that minimizes the overall cost, indicating the best matching between ground truth and predictions [16].The cost of matching considers the predicted class and the similarity of the predicted bounding boxes and ground truth boxes[15].

$$yi = \{bi, ci\}. \tag{2}$$

The vector bi denotes the positional coordinates of the ground truth bounding box's center, along with its height and width, which are normalized relative to the image dimensions and range from 0 to 1. Meanwhile Ci represents the probability assigned to the corresponding class label. On the other hand, the symbol ci represents the probability assigned to the corresponding class label for the ground truth bounding box. It indicates the likelihood or confidence of the object belonging to a specific class. The class labels typically represent different categories or classes such as "safety helmet," "no helmet," or other relevant labels in the context of safety helmet detection. Thus, Lmatch $(yi, \hat{}y(i))$ is[15]:

$$-1\ \{ci \neq \emptyset\}\ \hat{p}_{\sigma(i)}(c_i)\ +\ 1\ \{ci \neq \emptyset\}\mathcal{L}_{\text{box}}\ (b_1, \hat{b}_{\hat{\sigma}(i)}) \tag{3}$$

The key distinction between region proposal and anchor-based approaches is that DETR aims to achieve one-to-one matching for direct set prediction, eliminating the need for duplicates. The key distinction of the DeTr

Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

method is that it aims to achieve one-to-one matching for direct set prediction, eliminating the need for duplicate detections. Unlike region proposal and anchor-based approaches, DeTr directly predicts the set of objects without relying on intermediate region proposals or predefined anchor boxes. In the second step, the Hungarian loss is computed, which combines a negative log-likelihood for class prediction and a subsequently defined box loss[15].

$$\mathcal{L}_{Hungarian}(y_i, \hat{y} = \sum_{i=1}^{N}[-log\ \hat{p}_{\sigma(i)(c_i)} + 1\{ci \neq \emptyset\}\ \mathcal{L}_{Box}(b_1, \hat{b}_{\hat{\sigma}(i)})] \quad (4)$$

The optimal assignment is denoted as ^σ, calculated at the initial step. To solve the class imbalance problem, the log-probability term is reduced by a factor of 10 when Ci is the empty set. The loss function for the bounding boxes is a combination of the $\mathcal{L}_i$ loss and the generalized IOU loss, where the two losses are linearly combined[15].

$$\lambda_{iou}\mathcal{L}_{iou}(b_1, \hat{b}_{\hat{\sigma}(i)}) + \lambda_{L1}||\ b_1 - \hat{b}_{\hat{\sigma}(i)}|| \quad (5)$$

Both the IOU loss and the $\mathcal{L}_i$ loss capture different aspects of the bounding box prediction accuracy. The IOU loss focuses on the overlap between the predicted and ground truth boxes, while the $\mathcal{L}_i$ loss measures the absolute differences in the box coordinates. By combining these two losses in a linear manner, the DeTr method aims to balance their contributions and account for both localization accuracy and similarity with the ground truth. Using only the $\mathcal{L}_i$ loss may lead to inconsistent scales for small and large bounding boxes, even when their relative errors are similar. By including the IOU loss, which considers the spatial intersection and union of the boxes, the DeTr method addresses this issue and provides a more robust loss function for bounding box regression[15]. When the testing phase is critical in evaluating the performance and effectiveness of the model being trained. In this study, the testing scenario was to evaluate the data set used during the training stage. This dataset is specifically focused on safety helmets. This means that the testing phase aims to evaluate the model's ability to accurately detect and classify safety helmets in different images. Through testing, the researchers can measure the performance of the resulting model.

2.4 Evaluation

After successfully training the model using the training data, the model is then tested using the DeTr model on the designed system. Performance testing involves measuring the intersection over the union (IoU). This metric is generally used to determine positive samples by calculating the overlap between the anchor box (proposed bounding box) and the ground truth (GT) bounding box. The IoU is calculated by dividing the area of the junction between the two bounding boxes by the area of their union. A high IoU indicates a better match between the predicted bounding box and the ground truth, indicating a positive detection. By evaluating the IoU, researchers can assess the accuracy and quality of model predictions and help determine successful positive samples [17]. IoU is a method of measuring the accuracy of object detection in a data set[18]. Accuracy threshold should be selected when using IoU as a metric[19].There are two commonly used IoU thresholds, which are 0.5 and 0.75[1]. IoU requires two elements, the bounding box area of the ground truth area, which is the real area, and the intersection and joint detection area. In conclusion from the previous information, IoU (Intersection over Union) is a method used to measure the accuracy of object detection in a data set. In object detection evaluation, IoU compares the intersection area between the detection prediction and the actual object area to the total area of both elements. Two commonly used IoU thresholds are 0.5 and 0.75, where an IoU above these thresholds indicates more accurate detection. The IoU gives an indication of how well the object detection model can map the object precisely. The selection of an appropriate IoU threshold is crucial in determining the acceptable detection criteria. The IoU can be calculated using equation (1)[20].

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \quad (6)$$

By comparing the IOU to a certain threshold *t*, we can classify the detection as true or false. Detection is considered correct if IoU ≥ *t*. If IoU < *t* then the detection is considered false[21]. Thus, IoU (Intersection over Union) in this study is used as an evaluation metric to distinguish correct or incorrect detections. By comparing the IoU value with a certain threshold, t, the detection can be classified as correct or incorrect. If IoU ≥ t, the detection is considered correct, which indicates a good match between the prediction box and the ground truth. However, if IoU < t, the detection is considered false, indicating that the prediction does not match the ground truth. In conclusion, the use of IoU with a threshold of t allows researchers to perform an objective assessment of the quality of detection performed by the model, assisting in determining the accuracy and success rate of detection in a system.

3. RESULT AND DISCUSSION

In this case study, a DeTr model is used to train and test safety helmet detection. A model trained using training data outputs a detection transformer model that is used to detect test data.

Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

3.1 Dataset

The dataset used in this study amounted to 5000 images in image format jpg data type, and derived from Kaggle "Safety Helmet Detection" [22]. This dataset is divided into 70% training data, 20% validation data and 10% testing data. Table 1 shows the amount of total datasets. Figure 6 is an example dataset of Safety Helmet Detection used for training. The dataset consists of various conditions.

Table 1. Dataset

Data	Images Total
Training	3500
Validation	1000
Testing	500

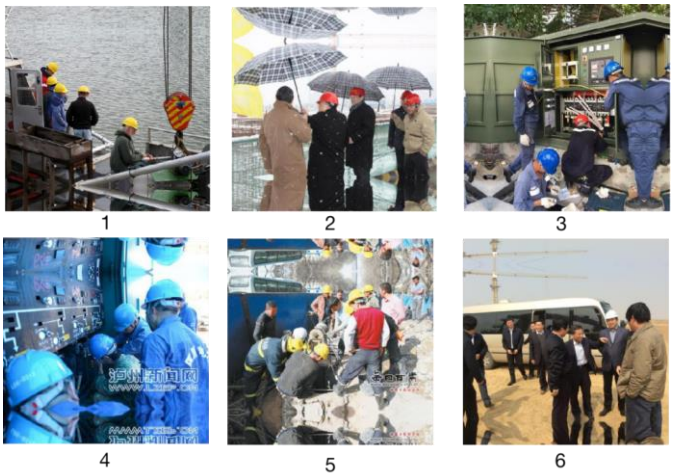

Figure 6. Example of dataset

3.2 Training Results

The training result is the result of the training performed on the Detection Transformer model using the prepared dataset. Before training, the training data is resized to 224x224. This process is done to make training with the Detection Transformer (DeTr) model faster and easier. Although the size of the training data image has been changed to 224 x 224, the results of the pre-trained detection transformer model still achieve good results with the hyperparameters shown in Table 2.

Table 2. Hyperparameter

Max Epochs	Gradient Clip Val	Accumulate Batches	Log Every Steps
50	0.1	8	5

Gradient clip val is to define minimum and maximum clip values. If the gradient exceeds a certain threshold, it will be clipped to the threshold, if the gradient is below the lower bound, the gradient is also clipped to the lower threshold. Gradient accumulation is a technique where you can train on bigger batch sizes than your machine would normally be able to fit into memory and we use the minimum of max epoch which is 50, and it generates the log every 5 steps. After the hyperparameters are set, the training results show that the obtained results are generally good, as their mAP and mAR values are greater than 0.5. The results training of average precision can be seen in Table 3 and the result training of average recall can be seen in Table 4.

Table 3. Result of Average Precision

IoU	Area	Max Detections	AP
0.50:0.95	All	100	0.539
0.50	All	100	0.889
0.75	All	100	0.590
0.50:0.95	Small	100	0.311
0.50:0.95	Medium	100	0.597
0.50:0.95	Large	100	0.728

Table 4. Result of Average Recall

IoU	Area	Max Detections	AR
0.50:0.95	All	1	0.141

Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

IoU	Area	Max Detections	AR
0.50:0.95	All	10	0.568
0.50:0.95	All	100	0.613
0.50:0.95	Small	100	0.430
0.50:0.95	Medium	100	0.668
0.50:0.95	Large	100	0.790

Table 3 and 4 summarizes the IoU metric, average precision (AP), and average recall (AR) values for different IoU thresholds, areas, and maximum detections. Values reported are the average precision and recall at intersections over union (IoU) thresholds between 0.5 and 0.95, with a maximum detection rate of 100 being calculated. Precision is calculated based on the number of positive detections (crossing the IoU threshold), then the precision values that exceed each threshold are averaged to get the final value. We calculated the average precision and average recall over different areas (small, medium, and large). The small area is defined the small object less than 32x32 pixel scale, the medium area is defined object between 32x32 and 96x96 pixel scale, and the large area defined object larger than 96x32 pixel scale. Testing the Average Precision and Average Recall metrics was carried out 6 times due to testing the IoU steps and in different areas.

3.3 Testing Results

At this step, testing is carried out using a trained model detection transformer that has been previously trained. Testing is done by detecting test data totaling 500 images. Of the 500 test data results, 10 well-perceived image samples were used, the results are shown in Figure 7

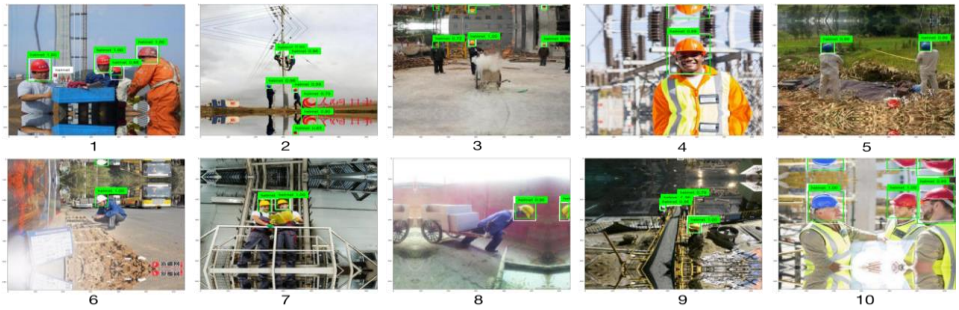

Figure 7. Sample of Correct

Based on the detection results using the Detection Transformer (DeTr) method in Figure 7, the results show that the system can detect safety helmets in the test data properly. Average IoU value in the sample shows that the system can detect properly can be seen in Table 5 below.

Table 5. Testing Results from Figure 7

	MIoU
Result	0.50

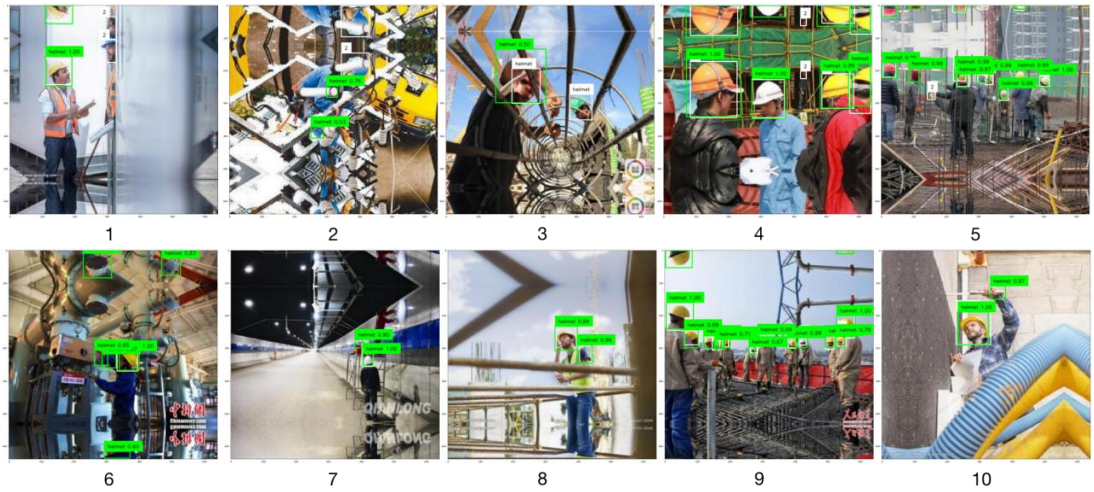

Figure 8. Sample of Incorrect

In Figure 8 above, it can be seen that the system has incorrect detection, some helmets are not detected because some objects are occluded by other objects as shown in results number 1 to number 3, then there are blurry images and small object, making it difficult for the system to detect helmet parts as shown in results number 4 and

Journal of Information System Research (JOSH)
Volume 4, No. 4, Juli 2023, pp 1316−1323
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v4i4.3852

number 5. As for results number 6 to number 10, the system has an incorrect in detecting objects that are assumed to be safety helmets.

4. CONCLUSION

Based on the test results, we built the system using Detection Transformer (DeTr) that can detect safety helmets in various conditions in the field project. Our study uses 500 testing data and tested with DeTr and can detect well. However, the system has incorrect detection such as the helmet is not detected because it is occluded between other objects, then there are blurry images and small object, making it difficult for the system to detect helmet parts, the system has a defect in detecting objects that assume that the object is a safety helmet. It can be concluded that the system of Detection Transformer can detect safety helmets but still has system defects with an average IoU value of 0.5.

REFERENCES

[1] T. A. A. H. Kusuma, K. Usman, and S. Saidah, "People Counting For Public Transportations Using You Only Look Once Method," Jurnal Teknik Informatika (Jutif), vol. 2, no. 1, pp. 57–66, Feb. 2021, doi: 10.20884/1.jutif.2021.2.2.77.
[2] M. Zhong and F. Meng, "A YOLOv3-based non-helmet-use detection for seafarer safety aboard merchant ships," in Journal of Physics: Conference Series, Institute of Physics Publishing, Nov. 2019. doi: 10.1088/1742-6596/1325/1/012096.
[3] H. Wang, Z. Hu, Y. Guo, Z. Yang, F. Zhou, and P. Xu, "A real-time safety helmet wearing detection approach based on csyolov3," Applied Sciences (Switzerland), vol. 10, no. 19, pp. 1–14, 2020, doi: 10.3390/app10196732.
[4] A. Hayat and F. Morgado-Dias, "Deep Learning-Based Automatic Safety Helmet Detection System for Construction Safety," Applied Sciences (Switzerland), vol. 12, no. 16, 2022, doi: 10.3390/app12168268.
[5] M. Lin et al., "DETR for Crowd Pedestrian Detection," 2020, [Online]. Available: http://arxiv.org/abs/2012.06785
[6] R. M. Mailoa and L. W. Santoso, "Deteksi Rompi dan Helm Keselamatan Menggunakan Metode YOLO dan CNN," Jurnal Infra, vol. 10, no. 2, pp. 56–62, 2022.
[7] S. B. Setyawan, W. Pribadi, H. Arrosida, and E. P. Nugroho, "Sistem Deteksi Pengendara Sepeda Motor Tanpa Helmdan Kelebihan Penumpang pada Dengan Menggunakan YOLO V3," Seminar Nasional Terapan Riset Inovatif, vol. 7,no. 1, pp. 430–438, 2021.
[8] R, M. "Survey on image preprocessing techniques to improve OCR accuracy," medium.com, 2021. [Online] Available: https://medium.com/technovators/survey-on-image-preprocessing-techniques-to-improve-ocr-accuracy-616ddb931b76.
[9] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani, "Rethinking Transformer-based Set Prediction for Object Detection," Oct. 2021, doi: https://doi.org/10.1109/iccv48922.2021.00359.
[10] Dedhia, P.R "Faster R-CNN: A step towards real-time object detection," medium.com, 2020. [Online] Available: https://towardsdatascience.com/faster-r-cnn-a-step-towards-real-time-object-detection-98c186732a69.
[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," May 2020, [Online]. Available: http://arxiv.org/abs/2005.12872
[12] Dai Zhigang, B. Cai, Y. Lin, and J. Chen, "UP-DETR: Unsupervised Pre-training for Object Detection with Transformers," Computer Vision and Pattern Recognition, Jun. 2021, doi: https://doi.org/10.1109/cvpr46437.2021.00165.
[13] Y. Liu, B. Schiele, A. Vedaldi, and C. Rupprecht, "Continual Detection Transformer for Incremental Object Detection," pp. 23799–23808, 2023.
[14] Hila Chefer, S. Gur, and L. Wolf, "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers," Oct. 2021, doi: https://doi.org/10.1109/iccv48922.2021.00045.
[15] S.-H. Tsang, "Review — DETR: End-to-End Object Detection with Transformers," towardsdatascience.com, 2022. [Online]. Available: https://sh-tsang.medium.com/review-detr-end-to-end-object-detection-with-transformers-c64977be4b8e.
[16] Yu, G., Xiang, N. and Pan, C. "Pedestrian detection in crowded scenes based on Cascade R-CNN" 2022 8th International Conference on Computer Technology Applications. doi:10.1145/3543712.3543720.
[17] K. Kim and H. S. Lee, "Probabilistic Anchor Assignment with IoU Prediction for Object Detection," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.08103
[18] W. Rahmaniar and A. Hernawan, "Real-time human detection using deep learning on embedded platforms: A review," Journal of Robotics and Control (JRC), vol. 2, no. 6. Department of Agribusiness, Universitas Muhammadiyah Yogyakarta, pp. 462-468Y, Nov. 01, 2021. doi: 10.18196/jrc.26123.
[19] H. Rezatofighi et al., "Generalized intersection over union: A metric and a loss for bounding box regression," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. doi:10.1109/cvpr.2019.00075
[20] G. Zhang, L. Ge, Y. Yang, Y. Liu, and K. Sun, "Fused Confidence for Scene Text Detection via Intersection-over-Union," Oct. 2019, doi: https://doi.org/10.1109/icct46805.2019.8947307.
[21] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Jul. 2020, doi: https://doi.org/10.1109/iwssip48289.2020.9145130.
[22] Larxel, "Safety helmet detection," Kaggle.com, 2020. [Online]. Available at: https://www.kaggle.com/datasets/andrewmvd/hard-hat-detection