

Deep learning-based localisation of combine harvester components in thermal images

Hanna Senke^{1,2}, Dennis Sprute², Ulrich B  ker³, and Holger Flatt²

¹ University of Applied Sciences and Arts Ostwestfalen-Lippe
Campusallee 12, 32657 Lemgo, Germany

² Fraunhofer IOSB, Industrial Automation Branch (IOSB-INA),
Campusallee 1, 32657 Lemgo, Germany

³ University of Applied Sciences and Arts Ostwestfalen-Lippe,
Institute Industrial IT (inIT),
Campusallee 6, 32657 Lemgo, Germany

Abstract It is crucial to identify defective machine components in production to ensure quality. Some components generate heat when defective, so automating the inspection process with a thermal imaging camera can provide qualitative measurements. This work aims to use computer vision methods to locate these components in thermal images. Since there is currently no comparison of object detection and semantic segmentation algorithms for this use case, this study compares different architectures with the goal of localising these components for further defect inspection. Moreover, as there are currently no datasets for this use case, this study contributes a novel annotated dataset of thermal images of combine harvester components. The different algorithms are evaluated based on the quality of their predictions and their suitability for further defect inspection. As semantic segmentation and object detection cannot be directly compared with each other, custom weighted metrics are used. The architectures evaluated include RetinaNet, YOLOV8 Detector, DeepLabV3+, and SegFormer. Based on the experimental results, semantic segmentation outperforms object detection regarding the use case, and the SegFormer architecture achieves the best results with a weighted MeanIOU of 0.853.

Keywords Thermal images, object localisation, deep learning architectures, industrial quality assurance

1 Introduction

Identifying defective components in production is crucial for quality management. Some components generate heat when defective and can be identified by this. Currently, this is either done manually or not done at all. An automatic inspection using a thermal imaging camera that captures temperature in a 2D image could enable objective and reproducible measurements, improving quality and supporting workers. To achieve this, the location of each component in the thermal image must be known. A naive approach is to use fixed areas. However, in modern production lines, there are often different machine variants with changing layouts, and some components can be very close to each other or even overlap. Thus, this simple approach does not provide the accuracy needed to evaluate components separately. To address this issue, the components have to be localised in each image individually based on computer vision algorithms, such as those from the fields of object detection and semantic segmentation.

Therefore, in this work, different object detection and semantic segmentation architectures are compared in an industrial production use case, specifically the localisation of combine harvester components during assembly as illustrated in Fig. 1. This use case is chosen due to the high number of product variants and component layouts. A main contribution of this work is a novel annotated dataset of thermal images of combine harvester components intended for object detection and semantic segmentation tasks. Moreover, this work provides a compre-

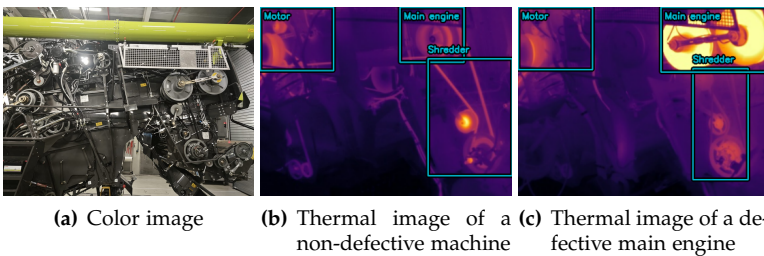


Figure 1: Images of combine harvesters in side view in different variants with the relevant components motor, main engine and shredder.

hensive performance evaluation of different object detection and semantic segmentation architectures on this novel dataset with the objective of localising the components for further defect inspection.

The remainder of this paper is structured as follows: First, existing approaches related to the topic and the architectures and backbones used for this study are presented. Then, the image acquisition and dataset generation process, the custom metric and the experimental setup are covered. Finally, the results of the comparison are discussed.

2 Related Work

There are already concepts for defect detection on thermal images [1], [2], [3]. However, most of these approaches localise the defect based on thermal information instead of localising the objects first. To ensure that each component is inspected separately, it is necessary to localise the components first. For object localisation on thermal images using object detection or semantic segmentation, there are already existing approaches, e.g. Mukherjee et al. [4] compare different versions of YOLO to localise humans and objects in disaster scenes. Ulhaq et al. [5] optimize YOLO to detect small objects for animal detection in thermal images. Moreover, Ippalapally et al. [6] detect objects for autonomous vehicles in the FLIR dataset, and Li et al. [7] propose a new architecture for semantic segmentation on thermal images. However, none of these approaches match this specific use case involving machine components of combine harvesters, which is especially challenging due the wide range of variants and component layouts.

There are also approaches that localise objects with the intention of further inspection. For example, Gong et al. [8] use YOLO as a base to localise electrical equipment and detect rotation to create a fitting area. Madura Meenakshi et al. [9] localise the eye region using YOLOV2 on infrared thermal images, while Kakileti et al. [10] use convolutional and deconvolutional neural networks on greyscale thermal images to segment areas for breast cancer detection. However, these works do not compare different object detection or semantic segmentation methods on thermal images with a focus on a subsequent inspection, which is necessary to select an optimal neural network architecture for component localisation.

3 Architectures and Backbones

To compare different object detection and semantic segmentation methods, four state-of-the-art neural network architectures are selected. SegFormer [11] is a transformer-based architecture for semantic segmentation that uses mix transformer (MiT) backbones. It can be compared to the DeepLabV3+ [12] architecture, which is also designed for semantic segmentation. DeepLabV3+ is a deep convolutional neural network (DCNN) and will be evaluated with common backbones, namely MobileNetV3, EfficientNetV2, ResNet, ResNetV2, DenseNet, and the YOLOV8 backbone. YOLO models are commonly used in object detection applications. For this study, the YOLOV8 Detector [13] is used and combined with the YOLOV8 backbones. It will be compared with RetinaNet [14], a popular one-stage object detection architecture that uses the same backbones as DeepLabV3+.

4 Image Acquisition, Preprocessing and Dataset

For the dataset of combine harvester components, thermal images were collected over 49 production days. The thermal camera captured an image every 10 to 20 seconds with a resolution of 382 by 288 pixels. Per day, there are five to ten measurement cycles, with each cycle testing one combine harvester. The data contains both defective and non-defective machines, with non-defective machines being more prevalent. The relevant components captured are the motor in the top left corner, the main engine in the top right and the shredder on the right side as illustrated Fig. 2. The temperature ranges from 30 °C to 60 °C for non-defective combine harvesters (see Fig. 2(a)) and from 30 °C to 125 °C for defective machines (see Fig. 2(b)). Due to the large number of variants, there are also machines without a shredder as depicted in Fig. 2(c). In total, this dataset comprises 19 different machine variants.

The acquired thermal images are first converted to RGB images to utilize common neural network architectures designed for colour images and pre-trained weights of large-scale image datasets. The temperature is clipped at 55 °C to account for inconsistent colouring due to higher temperatures in defective machines. To create the dataset, the measured data is split into measurement cycles. Then, three images per

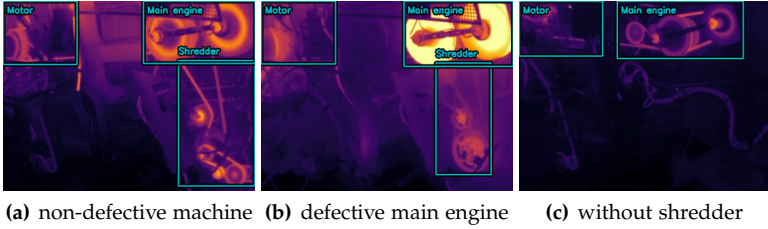


Figure 2: Thermal images of different combine harvester variants.

cycle with a temperature above 45°C are randomly chosen. This ensures that each machine variant is present in the dataset and different temperatures are represented. As there are significantly fewer images of defective combine harvesters, images with a temperature of 70°C or above are added to the dataset. Segmentation masks and bounding boxes are then manually added to the data. The final dataset consists of 1200 images, including 253 images of defective machines and 69 images of machines with only two components. This dataset is split into 720 training images and 240 validation and test images each.

The thermal images of defective combine harvesters have some differences in colouring and contrast. For the machines with two components, there is only a small number of images. To identify weaknesses in the models and highlight missing training data, additional datasets are needed. Separate test datasets for defective machines (DM) and non-defective machines (NDM), each containing 250 images, and a test dataset for machines with two components (TCM), containing 69 images, are created from images of the original dataset.

5 Custom Metric

For comparison of the performance of object detection models among each other and semantic segmentation models among each other, the Mean Intersection over Union (MeanIOU) is used. However, for object detection, MeanIOU does not provide a direct comparison for our specific use case, as the bounding boxes cannot accurately represent the components. To address this issue, the ground truth segmentation

masks are used for evaluation. However, bounding box predictions naturally include pixels that do not belong to the component. To ensure a fair comparison, non-heat generating parts of the machine, which do not pose a problem for defect inspection, need to be weighted differently compared to heat-generating parts.

For this purpose, a temperature threshold value $\tau = 70^\circ\text{C}$ is defined, which is specific to the components in this use case. The number of false negative pixels is denoted as FN . The false positive pixels are split into two groups using the temperature threshold τ . The number of pixels falsely classified as belonging to the component and with a temperature above the threshold τ is denoted as $FP_{t \geq \tau}$. The number of pixels falsely classified as belonging to the component with a temperature below the threshold τ is denoted as $FP_{t < \tau}$. Each group is given a separate weight. For one component, the weighted absolute error is defined as follows:

$$wAE = \lambda_1 \cdot FP_{t < \tau} + \lambda_2 \cdot FP_{t \geq \tau} + \lambda_3 \cdot FN \quad (1)$$

As mentioned before, non-heat-generating parts of the machine do not pose a problem but are often included in the prediction of object detection models. Therefore, λ_1 should be much smaller than the other weights. For this study, the false positives under the temperature threshold will be weighted with $\lambda_1 = 0.1$. Since the false positives over the temperature threshold τ and the false negatives influence the results of the final defect inspection, they will be weighted with $\lambda_2 = 1$ and $\lambda_3 = 1$.

The intersection over union (IOU) is a common metric to evaluate object detection and semantic segmentation, but it is not well suited for comparing predicted bounding boxes with ground truth segmentation masks. As the wAE evaluates the FP and FN , a weighted union wU can be calculated as the sum of the intersection and the wAE . For the weighted MeanIOU, the arithmetic mean over all components is calculated, with n representing the number of components. We define the weighted IOU and the weighted MeanIOU as follows:

$$wIOU = \frac{I}{wU} = \frac{TP}{TP + wAE} \quad (2)$$

$$wMeanIOU = \frac{1}{n} \sum_{i=1}^n wIOU_i \quad (3)$$

The precision can be weighted using the same concept as the other metrics, grouping the FP based on their temperature. Therefore, it is defined as follows:

$$wPrecision = \frac{TP}{TP + \lambda_1 \cdot FP_{t < \tau} + \lambda_2 \cdot FP_{t \geq \tau}} \quad (4)$$

Since $\lambda_3 = 1$, the regular recall does not need to be modified. As with the previous metrics, the arithmetic mean over all components is calculated for recall and weighted precision.

6 Experimental Setup

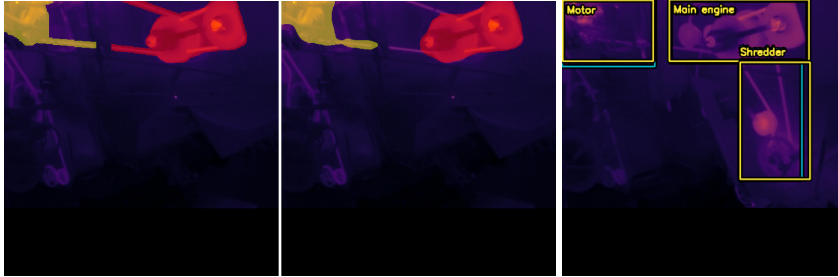
The experiments are conducted on a computer with Windows 11, equipped with a 12th Gen Intel Core i7 processor running at a base speed of 3.60 GHz, 32 GB of RAM, and an NVIDIA GeForce RTX 3060 graphics card with 12 GB of VRAM. The implementation is based on TensorFlow (2.16.1), Keras (3.0.5) and KerasCV (0.8.2).

All models are trained for 150 epochs using pre-trained weights from ImageNet or COCO dataset. At the end, the best weights based on the validation dataset are restored. The stochastic gradient descent (SGD) optimizer is used, with a global clipnorm of 10 and an exponential decay learning rate scheduler starting with a learning rate of 0.001. For each epoch, the model is trained on 360 images, which is half of the training dataset, and evaluated on the validation dataset. From each of the mentioned backbone types for DeepLabV3+ and RetinaNet in Sect. 3, one backbone, preferably of medium size, is selected. The YOLOV8 Detector and SegFormer are trained on all feasible backbones.

After training, the models are evaluated on the test dataset. Since the final goal is to classify objects as defective or non-defective, the most important metrics for the comparison are the weighted MeanIOU, weighted Recall, and weighted Precision. The MeanIOU is very suitable for comparing the performance of object detection and semantic segmentation models among each other.

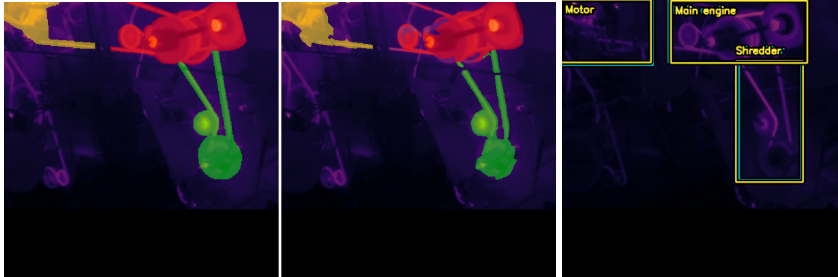
7 Results

The results of the three models with the best MeanIOU from each architecture are presented in Tab. 1, and an example prediction of the best model from each architecture is shown in Fig. 3. The DeepLabV3+



(a) SegFormer, MeanIOU = 0.805, wMeanIOU = 0.870

(b) RetinaNet,
MeanIOU = 0.915,
wMeanIOU = 0.56



(c) DeepLabV3+, MeanIOU = 0.805, wMeanIOU = 0.850

(d) YOLOV8 Detector,
MeanIOU = 0.944,
wMeanIOU = 0.589

Figure 3: Example predictions of the best models from each architecture. Ground truth on the left and prediction on the right for (a) and (c). Blue and yellow boxes represent the ground truth and model predictions, respectively, and the model's confidence score is visualized next to the bounding boxes for (b) and (d).

architecture achieves the best MeanIOU with the YOLOV8 M backbone and the second-best MeanIOU with the large MobileNetV3 backbone. For RetinaNet, the YOLOV8 M backbone achieves the best re-

sults, and the large MobileNetV3 reaches the second-best results. For the YOLOV8 Detector models, the medium-sized backbone achieves the best results.

Table 1: Results of the three best models from each architecture on the test dataset.

Architecture	Backbone	Pre-trained weights	Mean-IOU	wMean-IOU	Re-call	wPre-cision
SegFormer	MiT B0	ImageNet	0.8	0.853	0.861	0.988
DeepLabV3+	DenseNet169	ImageNet	0.755	0.804	0.812	0.971
	MobileNetV3 large	ImageNet	0.76	0.807	0.815	0.962
	YOLOV8 M	COCO	0.803	0.838	0.844	0.99
RetinaNet	DenseNet169	ImageNet	0.879	0.448	0.498	0.472
	MobileNetV3 large	ImageNet	0.901	0.513	0.579	0.534
	YOLOV8 M	COCO	0.914	0.682	0.766	0.704
YOLOV8 D.	YOLOV8 XL	COCO	0.936	0.631	0.707	0.651
	YOLOV8 XS	COCO	0.939	0.644	0.726	0.662
	YOLOV8 M	COCO	0.942	0.686	0.771	0.704

Compared by the MeanIOU, DeepLabV3+ with the YOLOV8 M backbone pre-trained on COCO performs the best for semantic segmentation. The second best performs SegFormer with the MIT-B0 backbone pre-trained on ImageNet. For the weighted MeanIOU, the SegFormer model performs better than the DeepLabV3 model. As the MeanIOU and the weighted MeanIOU are similar, there seems to be no significant difference between a transformer-based architecture and a DCNN for this use case. It is noticeable that colder parts of the component are not always detected, resulting in missing parts of predicted components. For object detection, compared by the MeanIOU, the YOLOV8 Detector performs better than the RetinaNet architecture. For the weighted MeanIOU both models perform similar. Both models use the YOLOV8 M backbone, pre-trained on the COCO dataset. The architectures sometimes miss components, especially in images with only two components.

Compared with the semantic segmentation models, the object detection models perform better for the MeanIOU. However, the semantic segmentation models perform better for the weighted MeanIOU.

Table 2: Models with the best MeanIOU from each architecture tested on the additional test datasets: non-defective machines (NDM), defective machines (DM) and machines with two components (TCM).

Architecture	Backbone	weighted MeanIOU			
		All	NDM	DM	TCM
RetinaNet	YOLOV8 M	0.682	0.65	0.529	0.793
YOLOV8 D.	YOLOV8 M	0.686	0.721	0.778	0.639
DeepLabV3+	YOLOV8 M	0.838	0.898	0.904	0.753
SegFormer	MiT B0	0.853	0.888	0.864	0.871

The semantic segmentation model with the best weighted MeanIOU, SegFormer with the MiT-B0 backbone, achieves a weighted MeanIOU of 0.853, while the best object detection model, the YOLOV8 Detector with the YOLOV8 M backbone, only reaches a weighted MeanIOU of 0.686. It is interesting to note that the best models from DeepLabV3+, YOLOV8 Detector, and RetinaNet all use the YOLOV8 M backbone pre-trained on the COCO dataset.

The results of the best model from each architecture on the additional test datasets can be seen in Tab. 2. For the additional test datasets, the YOLOV8 Detector and DeepLabV3+ perform worse on images with only two components than on images of defective machines. In contrast, RetinaNet and SegFormer perform better on images with two components than on images of defective machines.

8 Conclusions and Future Work

This study aimed to compare different object detection and semantic segmentation models with the objective of localising machine components in thermal images for further defect inspection. For this purpose, the specific use case of combine harvester components coming in a wide range of variants and layouts was selected. Based on the evaluation, semantic segmentation models provide the best results for the weighted MeanIOU, and the SegFormer architecture with MiT-B0 backbone achieves the best results. For the object detection architectures, the YOLOV8 M backbone performed best.

Additionally, the results show that the novel dataset presented challenges for the models. For images of defective machines, the colouring

differs from that of non-defective machines, resulting in less accurate predictions on these images. Additionally, images of machines with only two of the three components posed a problem. Both groups require better representation in the training dataset to address this issue. Overall, the dataset is quite small with 1200 images and could benefit from more data from additional measurement cycles. To overcome this problem, additional images could be artificially generated. This could be a promising area for future research, such as using stable diffusion techniques. For further investigation, it would be valuable to assess the performance of the models on data from non-harvesting machinery with different characteristics. Another potential study could explore the use of thermal data directly as a one-channel image with a modified architecture instead of converting it to an RGB image. Overall, this topic has a lot of potential for application in industrial production and quality assurance.

Acknowledgements

This research work is based on “Datenfabrik.NRW”, a flagship project by “KI.NRW”, funded by the Ministry for Economics, Innovation, Digitalisation and Energy of the State of North Rhine-Westphalia (MWIDE). We like to thank CLAAS for supporting the image data acquisition.

References

1. R. Ali and Y.-J. Cha, “Subsurface damage detection of a steel bridge using deep learning and uncooled micro-bolometer,” *Construction and Building Materials*, vol. 226, pp. 376–387, November 2019.
2. J. Hu, W. Xu, B. Gao, G. Y. Tian, Y. Wang, Y. Wu, Y. Yin, and J. Chen, “Pattern deep region learning for crack detection in thermography diagnosis system,” *Metals*, vol. 8, no. 8, p. 612, June 2018.
3. Y. Laib dit Leksir, M. Mansour, and A. Moussaoui, “Localization of thermal anomalies in electrical equipment using infrared thermography and support vector machine,” *Infrared Physics & Technology*, vol. 89, pp. 120–128, March 2018.
4. S. Mukherjee, O. Coudert, and C. Beard, “UNIMODAL: UAV-aided infrared imaging based object detection and localization for search and dis-

- aster recovery,” in *2022 Virtual IEEE International Symposium on Technologies for Homeland Security*, 2022, pp. 1–6.
5. A. Ulhaq, P. Adams, T. E. Cox, A. Khan, T. Low, and M. Paul, “Automated detection of animals in low-resolution airborne thermal imagery,” *Remote Sensing*, vol. 13, no. 16, p. 3276, June 2021.
 6. R. Ippalapally, S. H. Mudumba, M. Adkay, and N. V. H. R., “Object detection using thermal imaging,” in *2020 IEEE 17th India Council International Conference (INDICON)*, 2020, pp. 1–6.
 7. C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, “Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation,” *IEEE transactions on neural networks and learning systems*, vol. 32, no. 7, pp. 3069–3082, July 2021.
 8. X. Gong, Q. Yao, M. Wang, and Y. Lin, “A deep learning approach for oriented electrical equipment detection in thermal images,” *IEEE Access*, vol. 6, pp. 41 590–41 597, July 2018.
 9. R. Madura Meenakshi, N. Padmapriya, N. Venkateswaran, R. Ravikumar, and C. Ramya, “Localization of eye region in infrared thermal images using deep neural network,” in *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2021, pp. 446–450.
 10. S. T. Kakileti, G. Manjunath, and H. J. Madhu, “Cascaded CNN for view independent breast segmentation in thermal images,” *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2019, pp. 6294–6297, 2019.
 11. E. Xie *et al.*, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” May 2021. [Online]. Available: <http://arxiv.org/pdf/2105.15203>
 12. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” February 2018. [Online]. Available: <http://arxiv.org/pdf/1802.02611>
 13. “YOLOv8: A new state-of-the-art computer vision model,” 24.05.2024. [Online]. Available: <https://yolov8.com/>
 14. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 318–327, February 2020.