Laser flying-spot thermography: an open-access dataset for machine learning and deep learning

Kevin Helvig, Pauline Trouvé-Peloux, Ludovic Gavérina, Jean-Michel Roche, Baptiste Abeloos, Christophe Pradère

Laser flying-spot thermography: an open-access dataset for machine learning and deep learning

Kevin Helvig[a], P. Trouvé-Peloux[a], L. Gaverina[b], J-M. Roche[b], B. Abeloos[a], and C. Pradere[c]

[a]DTIS, ONERA - Université Paris-Saclay, F-91123, Palaiseau, France
[b]DMAS, ONERA - Université Paris-Saclay, F-92322 Châtillon, France
[c]EPSILON-ALCEN, esplanade des Arts et Métiers 33400 Talence, France

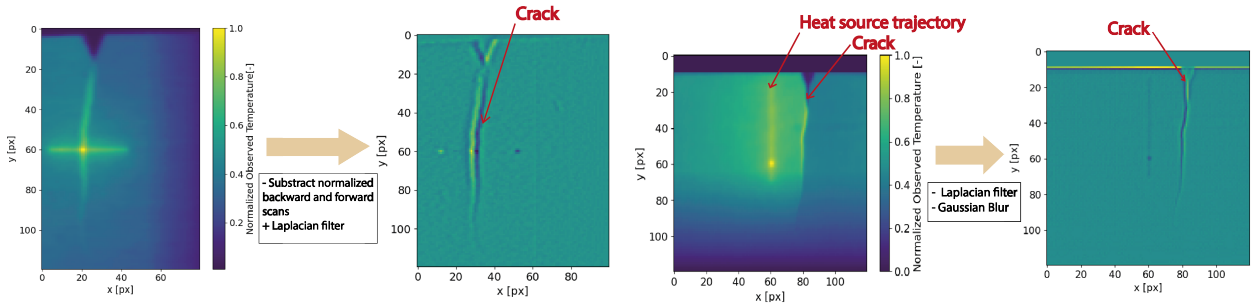
ABSTRACT

"Flying spot" laser infrared thermography (FST) is a non destructive testing technique able to detect small defects by scanning surfaces with a laser heat source. Defects, such as cracks on metallic parts, are revealed by the disturbance of heat propagation measured by an infrared camera. Deep learning approaches are now very efficient to automatically analyse and use contextual information from data, and can be used for crack detection. However, in the literature only few works deal with the use of deep learning for the crack detection in FST. Indeed obtaining a large amount of data from FST examinations can be expensive and time-consuming. We propose here to build a generic, open-access dataset of laser thermography for defect detection. This database can be used by the community to develop new crack detection methods that can be benchmarked on the same database, as well as for pretraining networks for similar application tasks. We also present results of state of the art detection networks trained with the proposed database. These models give a basis for future works. Dataset, called FLYD (FLYing spot thermography Dataset), will be available in : https://github.com/kevinhelvig/FLYD/.

Keywords: non-destructive testing, flying spot thermography, dataset, deep learning

1. INTRODUCTION

Review of literature. In order to inspect metallic aeronautical parts, FST can be used to provide local data for a precise detection and characterization of defects like surface cracks. The method is tolerant of low non-flatness of tested surface. This examination technique was originally developed for cracks detection in military aircraft parts in the end of the 1960s' [1]. In [2], Krapez made key theoretical work on this technique, studying the impact of experiment settings on crack detection accuracy. Many works have highlighted the interest of this technique in non-destructive testing, like for blade crack detection [3]. Regarding laser thermography, to the best of our knowledge, there are only few papers using deep learning for crack detection [4] and no open-access database in FST are available yet. Indeed, such FST experiment, requires expertise in thermal and laser fields, and expensive devices like infrared camera or power lasers. It also needs to conduct a substancial number of test with a large diversity of settings, which is time-consuming.



(a) (b)

Figure 1: (a) Left: reconstructed backward scan through defect. Right: Result of a Laplacian filter applied on the forward-backward scan difference [2]. (b) Reconstructed parallel thermal scan in defect region, with the Laplacian filtered image associated. All the images come from sample 4.

Krapez method is based on a "double pass" acquisition, i.e on the substraction between backward and forward scans crossing a crack, in order to attenuates edges and material heterogeneities. Crack is then detected using a high frequency filter, such as a Laplacian filter. However this method requires a time-consuming and hand-made adjustment of backward and forward thermal images, which is an issue for automation. Instead we propose to work in a "single pass" mode, performing only single scans. Heat source also scans parallel to the defect, to follow its penetration through the material. Difference between both Krapez and the proposed method are illustrated in Figure 1. Scans are converted into reconstructed thermal images which are normalized means of each frame of a recording. With the proposed "single pass" method, the data acquisition process is more simple, however the processing is then challenging, because crack has to be distinguised from edges and surface heterogeneities. Deep learning is a possible solution to this issue, able to learn contextual informations to perform detection. However with this technique large databases are necessary to train detection networks.

Contributions. The major contribution here is the gathering of amounts of data using an experimental set-up following the proposed single-pass FST approach. The dataset, named FLYD can be used to develop machine learning and deep learning approaches. The database contains raw-thermal recordings in "single pass" and a few "double-pass" scans Data are annotated to be used for binary classification between damaged and undamaged thermal images.A first benchmark for the classification task is proposed on the reconstructed single-pass thermal scans, testing various classic and state-of-the-art neural architectures.

2. DATASET: BENCH, SAMPLES AND DATA

This section describes the bench used for experimental data production and samples examined. Then we describe the proposed dataset for the binary classification task. This dataset is named FLYD-C (FLYD-Classification).

2.1 Flying-spot thermography bench

The FST bench of ONERA is illustrated in Figure 2. It uses a laser heat source with apower set to 2 W. Wavelength is 532 nm. A dichroic filter is added to reflect laser spot on the examined structure, and to allow IR to get to the MW-IR camera, sensitive between 3 and 5 µm. This wavelength band gives the possibility to study elevations of observed temperature of 10 degrees maximum, as induced by our heat source. The acquisition frequency is 100Hz, maximizing the number of frames collected compared to thermal specifications. The laser spot-size is chosen between two values: approximately 0.5 mm and 1.5 mm. The thermal spot covers a variable distance during a scan, and the scan velocity varies between two values: 0.5 mm/s and 1.5 mm/s.

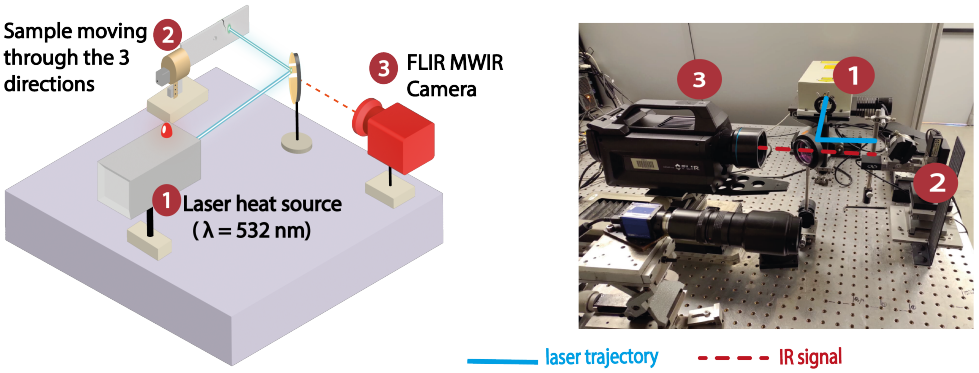


Figure 2: Illustration of the FST bench of ONERA.

2.2 Samples

The examined samples are ten metallic test specimens that were previously submitted to tensile test or fatigue stress test. Each of them presents a crack of various lengths, various openings or various orientations. Table 1 shows estimated crack length of each sample. It highlights the variability of length in our data amount. Crack opening varies also a lot between each sample: cracks of 4 to 6 and 9, 10, are easily distinguishable through eyes.

Defects of specimens 1, 3 and 8 are more difficult to see, while 2 and 7 are barely invisible through the naked eyes. A metallic specimen presenting a lare failure pattern was also examined and numbered sample 11: it gives examples without defects for our dataset. Each corner of this sample is scanned and named 11A to 11D.

Sample	1	2	3	4	5	6	7	8	9	10
Crack length estimation [mm]	6	8	7	13	9	15	11	8	17	23

Table 1: Samples with associated crack length.

Figure 3 illustrates the scanning protocol followed to perform data acquisition. Variations of velocity are included to increase the amounts of data and thermal properties presented in our dataset. The scan length varies, depending on starting position: scans A and C are 5 mm scans. Scans B are 8 mm scans. Scans D are 12 mm. Samples data are split between train-set and test-set in order to balance crack length in both set. Hence, samples 1 to 5, 8, 9, 10, and 11A, 11B, 11C give train data, while 6, 7 and 11D give test data.

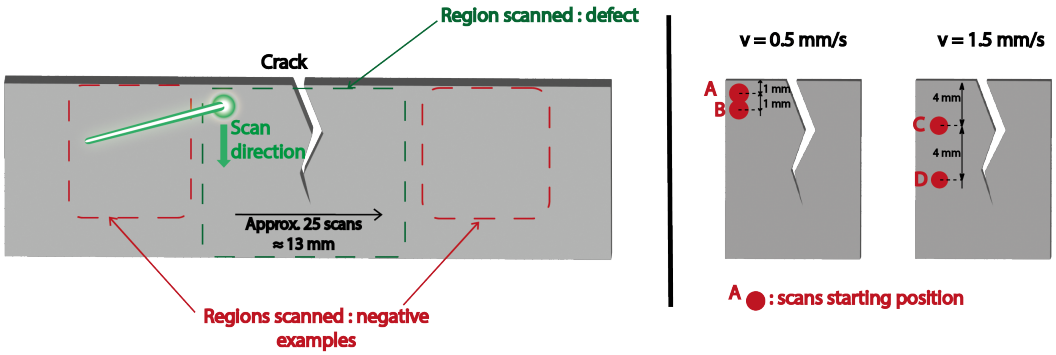
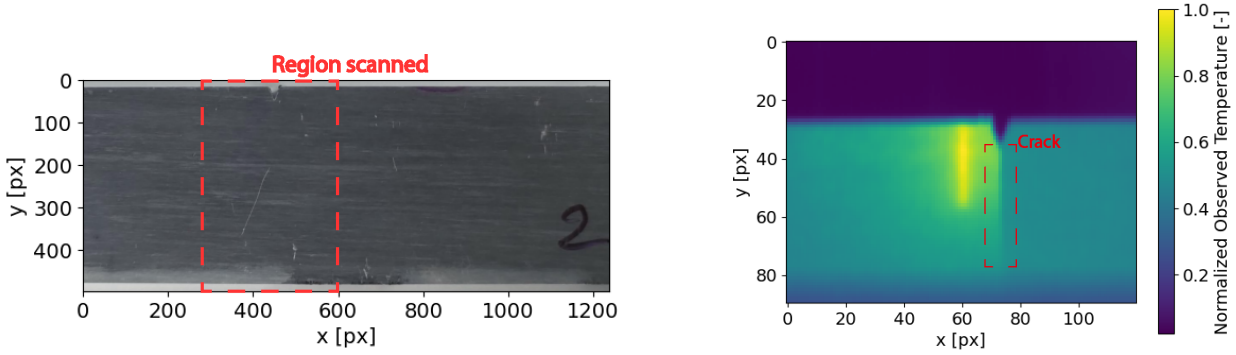


Figure 3: Experimental protocol followed in order to gather thermal recordings Same starting positions (given at right) have been applied for both defect and no-defect regions.

Figure 4 shows a visible image of sample 2, which presents a challenging defect which is barely invisible without optical means, and the associated reconstructed thermal scan. The relatively important penetration of the defect into core material is easy to distinguish in the IR spectrum. It promotes the interest of thermal techniques to proceed to fast surface examinations, in order to reduce false positives and missed detections during examination.



(a) (b)

Figure 4: (a) The visible spectrum acquisition of the defect region (sample 2). (b) Reconstructed thermal scan in defect region (sample 2). The crack is distinguishable in the IR spectrum.

2.3 Datasets: description

An amount of 1800 thermal scans is obtained by FST examination and converted into reconstructed thermal images. For binary classification images are labelled between uncracked and cracked classes. Data amount is

refined in order to have a balanced dataset. Dataset FLYD-C is splitted between 891 thermal images for training, and 286 for test set. An example from each class is given in Figure 5.

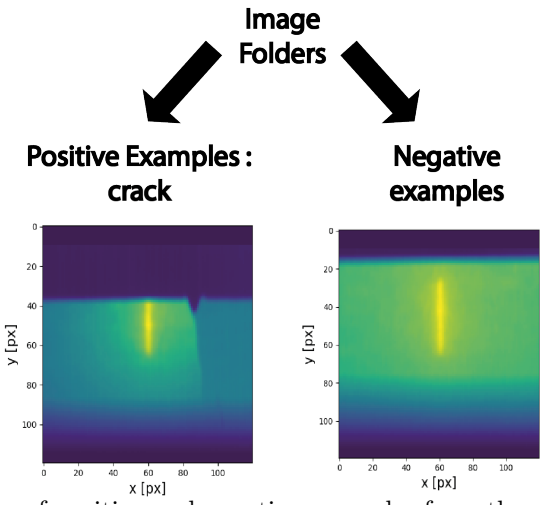


Figure 5: Examples of positive and negative examples from the classification dataset.

3. EXPERIMENTS

In this section we present a benchmark of performance of several neural architectures for binary classification task on our dataset. We provide in the following the training settings, metrics and architecture.

3.1 Classification task

Architectures. Several state of the art architectures for defect detection have been trained on this dataset. Our baseline of classical machine learning is a histogram-of-oriented-gradients (HOG) filter coupled with a support vector machine (SVM). This baseline is challenged by two convolution-based visual geometrics group (VGG) architectures: VGG 13 and VGG 16 [5]. These two neural networks are very traditionnal in the literature. Moreover, architectures using attention mechanism instead of convolutions are also trained and tested in our dataset. We assumed here that self-attention, giving ability to learn to hierarchize extracted information from different regions of thermal images, may help to detect defects by accomodating better to thermal phenomena [6]. Typical vision transformers architectures (ViT) have also been trained, such as ViT Base and ViT Large (respectively ViT-B and ViT-L in Table): the major difference between both is the number of attention stages, tending to increase both performance and data starvation [7]. Traditionnal vision transformers are more sensitive to data amounts, so we also compare performance with a recent modified and enhanced transformer architectures giving promising performance on similar tasks: CaiT (Class attention in image transformers) and Swin (shifted attention windows) [8; 9]. For a fair comparison between convolution-based and transformer-based architectures in modern deep learning, a ConvNext architecture, specifically built to challenge vision transformers, is trained and compared [10].

Method. For training ADAM optimizer has been used, with cross-entropy loss and for 200 epochs. Learning rate begins at 1e-4, with scheduling [11]. For each architecture, we initialize the weights with either random values or with pretrained values using the Imagenet database. Pytorch image models library (timm) gives access to pretrained architectures [12]. Data are split between train and test set as explained in section 2. Data augmentations, which are horizontal/vertical flips and random rotations, are implemented during training. We evaluate several metrics such as accuracy, F1, precision and recall are evaluated on test-set [13]. Precision and recall help to monitor missed detections and false alarms while F1 is a combination of both. Metrics are evaluated on test-set, and model with best accuracy is selected.

Results. Scores obtained by selected architectures are available in Table 2. The Table shows that all the deep learning based methods give generally better performance than the baseline using SVM, even only with

a pretrained model. Pretraining of the networks using ImageNet improves performance for a large majority of architectures, and especially for transformers architectures. Attention-based architectures seem to outperform the majority of convolution-based networks on our task, especially with pretrained models. However the ConvNext architecture is also close to transformers performance, mitigating this conclusion.

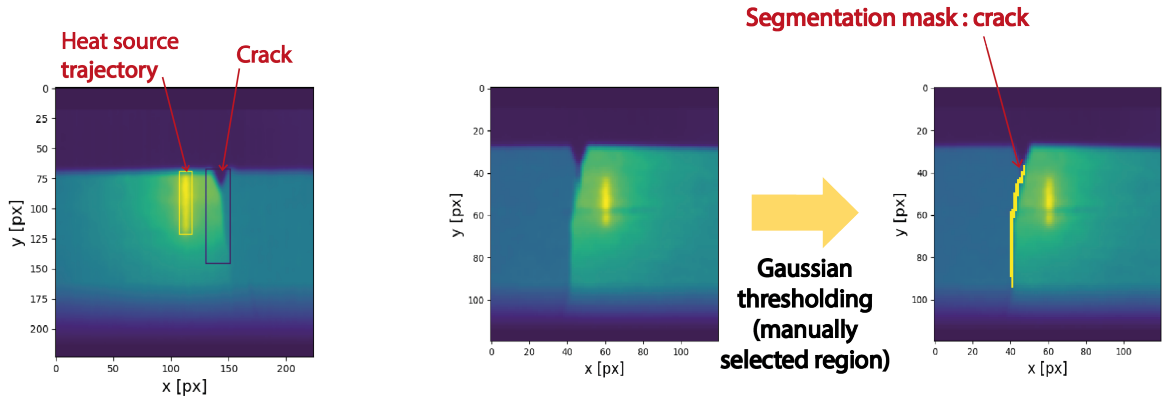
Architecture	Init. weights	Test-accuracy	F1	Precision	Recall
Baseline: HOG+SVM	-	79.1	0.79	0.78	0.79
VGG13	Random Init.	83.9	0.840	0.910	0.781
	Imagenet Init.	90.2	0.902	0.985	0.832
VGG16	Random Init.	75.5	0.713	0.977	0.561
	Imagenet Init.	81.1	0.795	0.963	0.677
ConvNext	Random Init.	-	-	-	-
	Imagenet Init.	98.6	0.990	0.987	0.994
ViT-B	Random Init.	86.7	0.881	0.850	0.916
	Imagenet Init.	98.6	0.987	0.975	0.999
ViT-L	Random Init.	84.3	0.862	0.819	0.910
	Imagenet Init.	99.0	0.990	0.981	1.00
Swin	Random Init.	-	-	-	-
	Imagenet Init.	98.9	0.990	0.987	0.993
CaiT	Random Init.	-	-	-	-
	Imagenet Init.	**98.9**	**0.990**	**0.981**	**0.999**

Table 2: Performance benchmark for the binary classification dataset.

4. FUTURE DATASET EXTENSIONS

Detection and segmentation tasks. In further works, our aim is to enrich our dataset with label for a detection and localisation task. This instance of our dataset will be called FLYD-D. Thermal images will be annotated using Labelstudio software to add ground-truth bounding boxes. It would be interesting to annotate images in order to locate two objects: trajectory of heat source trajectory and crack. Labels for heat source trajectory may be useful to develop deconvolution methods to identify and erase heat source directly from scan using learned approaches or not. Annotation format chosen will be MS-COCO 2017 [14].

It could also be interesting to provide a dataset for a segmentation task. To do so, defect will be thresholded using Otsu segmentation in a region close to the defect, on each images. Segmentation instance of our dataset will be named FLYD-S. Figure 6 shows an example of both tasks with associated annotations.



(a) (b)

Figure 6: (a) Detection and location task example: bounding boxes are displayed for heat source trajectory and defect. (b) At left: reconstructed thermal scan. At right: Reconstructed thermal scan with manual segmentation of crack in yellow.

Several state of the art detection and localisation architectures could be used on our dataset, such as a Faster-RCNN [15] as a baseline and the recent YOLO v5 architecture [16]. Deformable-DETR architecture could also be a challenger from transformers applied for detection task [17].

Adding data with more various FST settings and samples thermal properties. In our dataset the laser power is fixed, meaning that the signal to noise ratio of the data doesn't vary much in the dataset. It could be interesting to add thermal images with various various laser powers. It might help to make neural architectures more robusts to this parameter. Note that influence of experimental settings on thermal response of FST has been proposed in our preliminary study for different samples[18]. We will provide also data using "double-pass" examination procedure to provide quantitative comparison of crack detection between this approach and the proposed "single pass" approach. Finally, our dataset will be continously enriched by annotating samples with more diverse thermal properties, giving the possibility to study their influence on the crack detection performance.

5. CONCLUSION

This paper presents FLYD, a dataset for crack detection with laser infrared thermography, in single-pass examination mode. We provide a first benchmark to compare different machine learning and deep learning approaches on FLYD-C classification subset. Our tests illustrate that attention mechanism is relevant for active IR image processing. This publicly available dataset will hopefully provide a first basis for pretraining of neural architectures for similar tasks, such as other thermal examination techniques based on the use of a local heat source. This dataset will be continously improved with additionnal thermal images with more various experimental settings of the FST bench, samples with various thermal properties, as well as thermal images in the more traditionnal double-pass mode. Annotations for detection-location and segmentation tasks will also be added in further works.

References

[1] E. J. Kubiak, "Infrared Detection of Fatigue Cracks and Other Near-Surface Defects," *Applied Optics* **7**, pp. 1743–1747, Sept. 1968. Publisher: Optical Society of America.

[2] J.-C. Krapez, "Résolution spatiale de la caméra thermique à source volante," *International Journal of Thermal Sciences* **38**, pp. 769–779, Oct. 1999.

[3] N. W. Pech-May and M. Ziegler, "Detection of Surface Breaking Cracks Using Flying Line Laser Thermography: A Canny-Based Algorithm," *Engineering Proceedings* **8**, p. 22, Nov. 2021.

[4] W. Shi, Z. Ren, W. He, J. Hou, H. Xie, and S. Liu, "A technique combining laser spot thermography and neural network for surface crack detection in laser engineered net shaping," *Optics and Lasers in Engineering* **138**, p. 106431, Mar. 2021.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]* , Apr. 2015. arXiv: 1409.1556.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Dec. 2017. Number: arXiv:1706.03762 arXiv:1706.03762 [cs].

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," June 2021. arXiv:2010.11929 [cs].

[8] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with Image Transformers," Apr. 2021. Number: arXiv:2103.17239 arXiv:2103.17239 [cs].

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Aug. 2021. arXiv:2103.14030 [cs].

[10] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," Mar. 2022. arXiv:2201.03545 [cs].

[11] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv* , Dec. 2014.

[12] R. Wightman, N. Raw, A. Soare, A. Arora, C. Ha, C. Reich, F. Guan, J. Kaczmarzyk, mrT23, Mike, SeeFun, contrastive, M. Rizin, H. Kim, C. Kertész, D. Mehta, G. Cucurull, K. Singh, hankyul, Y. Tatsunami, A. Lavin, J. Zhuang, M. Hollemans, M. Rashad, S. Sameni, V. Shults, Lucain, X. Wang, Y. Kwon, and Y. Uchida, "rwightman/pytorch-image-models: v0.8.10dev0 Release," Feb. 2023.

[13] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv* , Oct. 2020.

[14] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," Feb. 2015. arXiv:1405.0312 [cs].

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jan. 2016. arXiv:1506.01497 [cs].

[16] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, Lorna, Yifu), C. Wong, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain, "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," Nov. 2022.

[17] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," Mar. 2021. arXiv:2010.04159 [cs].

[18] K. Helvig, L. Gaverina, P. Trouvé-Peloux, J.-M. Roche, B. Abeloos, C. Pradere, and G. Le Besnerais, "Towards deep learning fusion of flying spot thermography and visible inspection for surface cracks detection on metallic materials," in *The biannual Quantitative InfraRed Thermography (QIRT) 2022*, (Paris, France), July 2022.