Composite Image Generation Using Labeled Segments for Pattern-Rich Dataset without Unannotated Target

Kazuma Kano
Nagoya University
Nagoya, Japan
kazuma@ucl.nuee.nagoya-u.ac.jp

Yuki Mori
Nagoya University
Nagoya, Japan
ymori@ucl.nuee.nagoya-u.ac.jp

Keisuke Higashiura
Nagoya University
Nagoya, Japan
urachan@ucl.nuee.nagoya-u.ac.jp

Tahera Hossain
Nagoya University
Nagoya, Japan
tahera@ucl.nuee.nagoya-u.ac.jp

Shin Katayama
Nagoya University
Nagoya, Japan
shinsan@ucl.nuee.nagoya-u.ac.jp

Kenta Urano
Nagoya University
Nagoya, Japan
urano@nagoya-u.jp

Takuro Yonezawa
Nagoya University
Nagoya, Japan
takuro@nagoya-u.jp

Nobuo Kawaguchi
Nagoya University
Nagoya, Japan
kawaguti@nagoya-u.jp

Abstract

Although object detection technology using cameras offers potential for various applications, it incurs dataset creation costs to train new models where general-purpose models are ineffective, such as in industrial settings. We have previously developed a semi-automated annotation framework that employs optical flow and representation learning techniques to reduce human effort significantly. However, it was likely to cause unintended annotation omissions and mistakes compared to manual annotation. In this study, we propose a composite image generation approach to create omission-free and pattern-rich datasets. The proposed method synthesizes natural-looking images without unannotated targets by placing labeled foreground segments at their original positions on targetless background frames collected with the same fixed-point cameras. Evaluation with video footage in a logistics warehouse confirmed that improved dataset reliability led to higher model performance.

CCS Concepts

• **Computing methodologies** → **Image processing**; **Object detection**.

Keywords

Data augmentation, Image synthesis, Logistics warehouse, Optical flow, Representation learning

ACM Reference Format:
Kazuma Kano, Yuki Mori, Keisuke Higashiura, Tahera Hossain, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2024. Composite Image Generation Using Labeled Segments for Pattern-Rich Dataset without Unannotated Target. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '24), October 5–9, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3675094.3678447

UbiComp Companion '24, October 5–9, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1058-2/24/10
https://doi.org/10.1145/3675094.3678447

1 Introduction

Cameras are installed everywhere and used for various purposes, such as security and healthcare. With the recent aging population, they also hold promise for applications in work analysis for productivity improvement at industrial sites [17]. Object detection, a common task in computer vision, helps digitize valuable information regarding locations and status. Many current object detection technologies are based on deep learning and have achieved high accuracy. People can readily benefit from these technologies using publicly available pre-trained models, e.g., SSD [10], Faster R-CNN [12], DETR [1], and YOLO. Nevertheless, these general-purpose models may not be robust enough for specialized situations like industrial settings. For example, logistics warehouses have diverse items and unique equipment, such as hand pallets, which may be unfamiliar to the models. Additionally, although mounting wide-angle cameras at vantage points can prevent tall objects from screening the targets and cover extensive areas with fewer devices, views from right above or distorted are not generally anticipated. In this case, it is necessary to prepare datasets for fine-tuning and transfer learning, and the annotation costs hinder putting the systems into practice.

Various studies have been conducted to reduce the costs, such as effectively training models with less data and partially automating annotation tasks through computer assistance. We have previously developed a semi-automated annotation framework that extracts moving objects by optical flow, encodes them with representation learning, and groups similar ones by clustering [8]. It dramatically reduced human effort by setting bounding boxes automatically and labeling many objects together. However, it was prone to unintended annotation omissions compared to manually because it could not annotate objects stationary or not successfully grouped, which may impair the training processes.

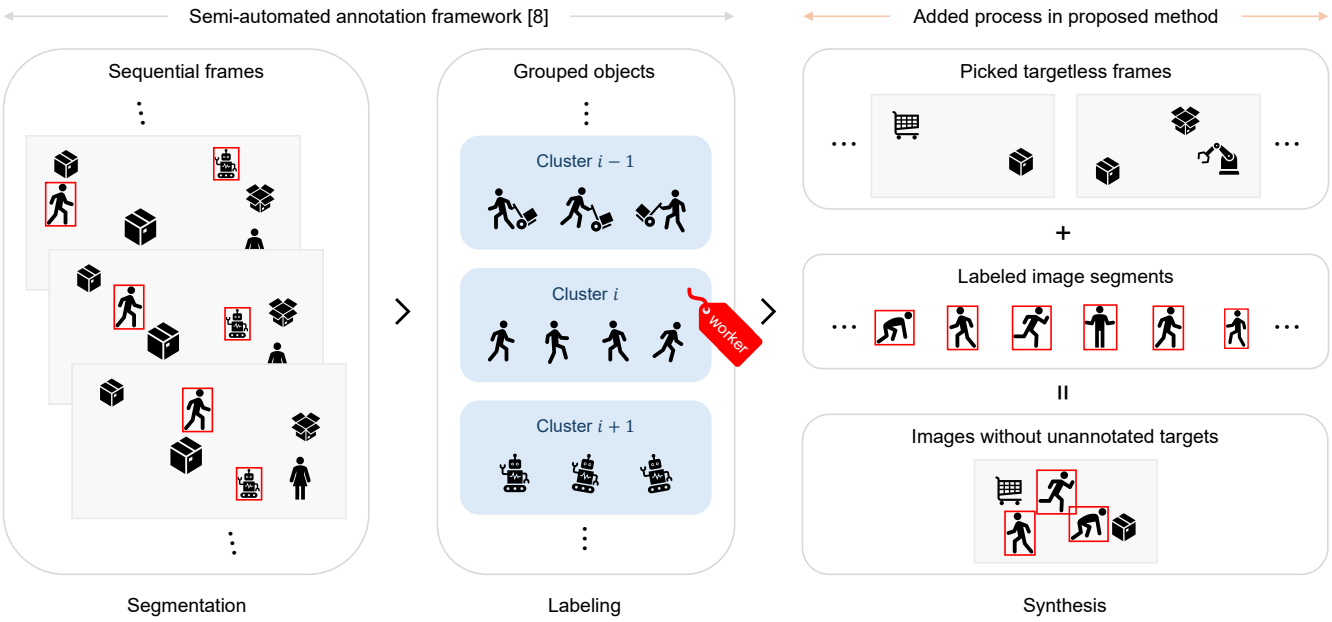

Figure 1: Composite Image Generation Using Labeled Image Segments.

Therefore, this study proposes a composite image generation approach to create omission-free and pattern-rich datasets, extending the framework with minor additional efforts. Figure 1 illustrates the concept of the proposed method. It picks frames without detection targets from candidate frames provided by optical flow. Then, it pastes image segments labeled with the framework onto the frames to synthesize fully annotated images. Here, we collect the frames and segments with the same fixed-point cameras and place the segments in their original positions and orientations to align the contexts of foregrounds and backgrounds. We created datasets with annotations of warehouse workers using video footage from wide-angle cameras on the ceiling and evaluated the performance of detection models trained with them. The result demonstrated that ridding datasets of unannotated targets improved the model performance by over 5 [%], and the proposed method achieved competitive accuracy with manual annotation at under a quarter of the cost.

2 Related Work

2.1 Active Learning

Previous studies have addressed the annotation cost problem in varied ways. Active learning is an approach where models initiatively select effective data points and prioritize annotating them. Yang et al. utilized similarity and uncertainty information provided by Fully Convolutional Network (FCN) to determine the most representative and unexplored data [16]. Yoo et al. attached a small parametric module to the network and predicted the losses for unlabeled inputs, i.e., how likely to go wrong [18]. Choi et al. used Mixture Density Network (MDN) for object detection to estimate aleatoric and epistemic uncertainty on localization and classification in a single forward pass [3]. These studies have enabled high accuracy with fewer data but still require considerable amounts, demanding further cost reductions.

2.2 Annotation Automation

Some studies reduced human effort by automating portions of annotation tasks. Lu et al. introduced self-supervised contrastive learning to train models partially with unannotated data [11]. They achieved high accuracy in lung nodule malignancy and attribute prediction with hundreds of samples and slight annotations. Elangovan et al. employed machine learning techniques to automate subtasks in the annotation process and created a dataset including over 10000 kitchen activities labeled with 24 attributes [6]. Nevertheless, these methods depend on specific domains and are difficult to apply directly to environments like multi-product warehouses. We have tackled generic semi-automated annotation for object detection of mobile classes, including humans, robots, and anything they carry [8]. It significantly reduced the costs for large-quantity annotation but had challenges of annotation omissions and mistakes.

2.3 Data Augmentation with Image Synthesis

Image synthesis is sometimes employed to simulate hard-to-obtain data or enrich on-hand data. Sakaridis et al. created foggy scene images by overlaying synthetic fog on real clear-weather scenes based on depth information [13]. On the other hand, Dwibedi et al. pasted object image segments onto scene images to generate augmented similar scenes, blending with several modes to make models ignore the pixel artifacts [5]. However, they assumed access to object images with modest backgrounds from multiple viewpoints; it is costly to collect those of practical targets, such as workers with various postures seen in the operations. In addition, they randomly placed the foreground segments on the background images, which

results in unnaturalness, especially when using distorted images from wide-angle cameras. Even though some studies estimated the semantic and geometric contexts with dedicated models to arrange the segments appropriately [4, 7], it could be difficult for complicated environments.

3 Methodology

3.1 System Overview

So far, we have developed a semi-automated annotation framework for mobile objects [8]. In this study, we apply image synthesis to the framework to eliminate unannotated targets and ensure data variety. Figure 2 outlines the procedure of the proposed method. The blue and orange steps stand for autonomous and human-involved, respectively. Step 1 segments objects based on the magnitude of the pixel motions computed by RAFT, a dense optical flow method with deep learning [14]. Step 2 encodes the segments into fixed-length vectors with a SimSiam model, a self-supervised representation learning method with Siamese-like network architecture [2]. We preliminary trained the model so that it encoded similar objects closely. Step 3 clusters the vectors by K-Means. Step 4 labels the grouped similar objects in bulk. Step 5 picks frames without target objects from candidate frames where RAFT did not detect motion. Step 6 synthesizes annotated images by placing the labeled segments on the targetless frames.

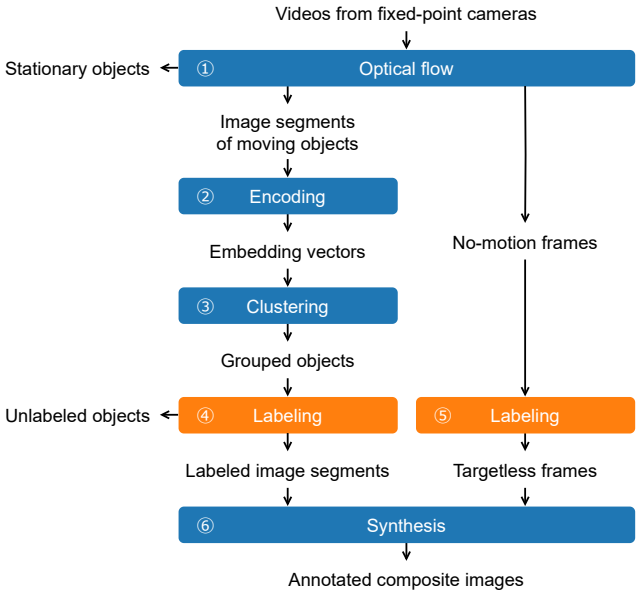

Figure 2: Procedure for Dataset Creation.

This study follows the previous work for steps 1 to 4 and appends steps 5 and 6. The previous method can fail to annotate in two ways: overlooking stationary objects at step 1 and not labeling misgrouped objects at step 4. The proposed method uses only labeled objects to create images without unannotated detection targets. It also allows for intentionally excluding poorly segmented objects, refining the quality of bounding boxes in the datasets. This method is applicable to object detection with fixed-point cameras for any mobile class. We defer the detailed explanation for steps 1 to 4 to the previous paper and focus on steps 5 and 6 in this chapter.

3.2 Targetless Frame Preparation

Background images without detection targets are necessary to ensure no unannotated targets in synthesized images. We utilize the motion information provided by RAFT to find the targetless images efficiently, assuming that the targets tend to move. The proposed method gathers targetless frames from randomly sampled frames where RAFT did not detect motion. Irrelevant objects in these frames are desirable for data diversity. In this step, humans need to check some frames for every camera, yet there are far lighter workloads than manual annotation. The more frames you prepare, the more diverse the generated images, but the higher the effort.

3.3 Composite Image Generation

We generate fully annotated composite images using the labeled image segments and targetless frames following Algorithm 1. The input consists of labeled segments $\boldsymbol{S}$, targetless frames $\boldsymbol{F}$, mean μ and standard deviation σ of the number of used segments per composite image, and iteration count C. The output is composite images $\boldsymbol{I}$. Each segment will be used C times on average. Equation (1) gives the expected proportion p of the segments used at least once, where N is the number of the segments.

$$p = 1 - \left(1 - \frac{\mu}{N}\right)^{Round\left(\frac{CN}{\mu}\right)} \tag{1}$$

Algorithm 1: Composite Image Generation

Data: Labeled image segments $\boldsymbol{S}$ of length N, targetless frames $\boldsymbol{F}$ of length M, mean μ and standard deviation σ of # of objects per image, and iteration count C.

Result: Composite images $\boldsymbol{I}$ of length $Round\left(\frac{CN}{\mu}\right)$.

```
begin
    for i ∈ I do
        set ⟵ {}
        num ⟵ Round(Gauss(μ, σ))
        while Len(set) < num do
            s ⟵ Choice(S)
            overlap ⟵ False
            for t ∈ set do
                if CheckOverlap(s, t) then
                    overlap ⟵ True
            if ¬overlap then
                set ⟵ set + {s}
        i ⟵ Choice(F)
        Sort(set, Centricity)
        for t ∈ set do
            i ⟵ Synthesize(i, t)
```

First, randomly select a labeled segment. Next, check whether it overlaps with any other segments already adopted for this time. Equation (2) determines it based on the area ratio of two bounding boxes' intersection to the smaller box, where $R_{overlap}$ represents the threshold ratio. We set $R_{overlap} = 0.25$ in this study.

$$CheckOverlap(s,t) = \frac{Area(s \cap t)}{\min(Area(s), Area(t))} > R_{overlap} \quad (2)$$

If the segment overlaps with no other segment, adopt it. Repeat these steps until sampling a random number of segments. Then, randomly select a targetless frame. Finally, synthesize an image by pasting the sampled segments onto the frame. Here, place the segments at their original positions and orientations in order of distances from the viewpoint to fit the foregrounds into the background context and simulate depth perception. We calculate distances from the center of the image instead because the cameras face right downward in our environment. Iterate the above steps for each camera to create a dataset.

4 Evaluation

4.1 Dataset Creation

We semi-automatically annotated workers in video footage of a receiving and shipping floor in a logistics warehouse. We used wide-angle cameras (H.View HV-800G2A5 [1]) fixed vertically downward on the ceiling. We recorded the videos in full HD, 8000 [kbps], and 5 [fps] and undistorted them with Double Sphere camera models [15]. Here are the labeling guidelines we followed to ensure the quality of annotations.

- Regard only objects composed of single workers as worker objects, ignoring ones comprising multiple workers or non-workers.
- Label only successfully segmented objects, excluding ones composed of partial bodies or including excessive margins.

As a result, we labeled 4066 objects as workers from 22000 moving objects over 13709 frames captured by 22 cameras. We also gathered five targetless frames, i.e., without workers, for each camera.

Then, we created datasets from these data in two ways: our previous method [8], i.e., just retrieving frames containing the labeled objects, and the proposed method, i.e., generating composite images from the labeled image segments and targetless frames. Table 1 presents the hyperparameters for composite image generation in the proposed method. Figure 3 displays data examples for the same three cameras; the left is retrieved frames by the previous method, and the right is composite images by the proposed method. The red solid rectangles represent actual annotations in the datasets, and the green dashed ones represent unannotated workers we marked for clarity. We can see that the previous method had some unannotated workers, whereas the proposed method eliminated them. Although the proposed method sometimes resulted in odd situations as seen in the top left of the upper image, the generated images were consistent with the perspective and mostly looked natural.

Furthermore, we created a dataset with almost the same number of annotations manually made by two experienced annotators. We also prepared another manually annotated dataset for the test. Table 2 summarizes the number of annotations and images in these datasets and the time spent on labeling and annotation by hand. The high number of annotations and images in the proposed method was due to the twofold augmentation of the original annotations. Meanwhile, the number of images in the previous semi-automated method was higher than in manual annotation owing to annotation omissions. Human work times in the proposed and previous method were under a quarter of manual annotation. Note that Table 2 does not include time to gather the targetless frames and execute the automatic processes.

[1] https://hviewsmart.com/products/h-view-colorcam-4k-bullet-ai-camera-with-color-night-vision-hv-800g2a5

Table 1: Hyperparameters for Composite Image Generation

Mean of # of objects per image μ	2
Standard deviation of # of objects per image σ	0.5
Iteration count C	2

Table 2: Data Quantities and Work Times

	# of Annots	# of Images	Time [min]
Proposed	8131	4066	94
Semi-automated [8]	4066	3786	94
Manual annotation	4062	1322	432
Test data	863	467	—

4.2 Model Training

To begin with, we split each dataset except the test into training and validation subsets in an approximate 8 : 2 ratio. The training subsets were randomly augmented in hue, saturation, brightness, translation, scale, shear, perspective, and flip. Then, we fine-tuned the medium-sized pre-trained model of YOLOv8 [9] with each dataset. This task was a single-class object detection, i.e., the models estimated bounding boxes enclosing workers. The batch size and maximum number of epochs were set to 128 and 300, respectively. We adopted the weights at the epochs with the best Average Precisions (AP) on validation subsets, following the default setting of the official implementation.

4.3 Results and Discussions

Figures 4 and 5 plot precision-recall curves and F1-confidence curves for the test data. Table 3 presents APs at that time. First of all, we will consider differences from the previous semi-automated method [8]. The proposed method outperformed the previous by over 5 [%] for all three AP metrics. Fully annotated data, a change from the previous, seemingly contributed to the model performance improvement. Figure 4 shows a significant drop in recall of the previous at around 0.85, which should be attributable to unannotated targets in the training data. The model likely learned to identify as workers only when particularly confident, increasing false negatives. Incidentally, these two methods exhibited similar trends in the high-precision range since these datasets originated from the same annotations.

(a) Semi-automated [8] (b) Proposed

Figure 3: Image Examples in Datasets.

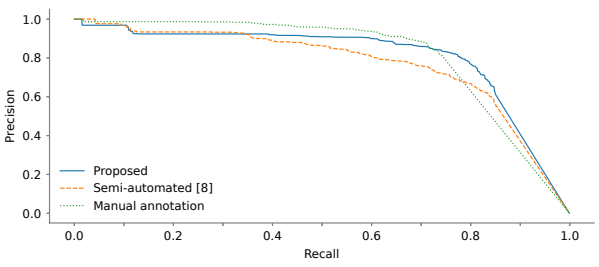

Figure 4: Precision-Recall Curves for Test Data.

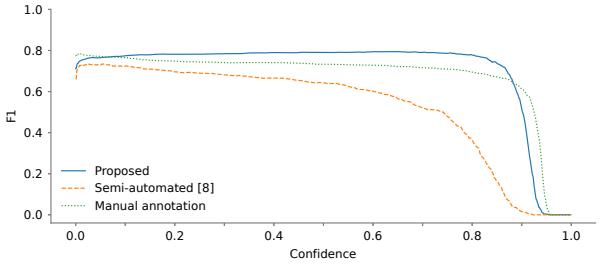

Figure 5: F1-Confidence Curves for Test Data.

Table 3: Average Precisions for Test Data

	AP50	AP75	AP50:95
Proposed	0.81	0.62	0.56
Semi-automated [8]	0.77	0.57	0.53
Manual annotation	0.82	0.64	0.57

Next, we will compare the proposed method with manual annotation and examine rooms for enhancement. The proposed method approached manual annotation in AP, especially excelling in recall and F1 score. The lack of negative data in the proposed method seems to have caused an increase in recall. This experiment focused on single-class detection, so the dataset of the proposed method, with its limited backgrounds, did not contain sufficient non-worker objects. The model may have learned to easily infer as workers, leading to higher recall at the expense of precision. In fact, precision hit the ceiling at around 0.9. The proposed method can be readily applied to multi-class detection. In this case, the classes act as negative data for each other, which could improve precision. Also, labeling non-target objects and synthesizing those segments will enrich background patterns, potentially enhancing precision.

5 Conclusion

This study tackled low-cost and reliable dataset creation for object detection, anticipating practical scenarios. We proposed a composite image generation approach using labeled image segments and targetless frames from fixed-point cameras. It enables the production of various realistic images by pasting the foreground segments at their original positions and orientations on the background frames. We extended a semi-automated annotation framework [8] and addressed the problem of unintended annotation omissions. Then, we created datasets of warehouse workers and evaluated the model performances trained with them. The result indicated that ridding the dataset of unannotated targets improved the model performance, and the proposed method was competitive to even manual annotation with less human effort. We also discussed challenges in the lack of negative data, which may lead to low precision. Applying multi-class and labeling non-target objects are possible solutions for further improvement. This approach to generating composite images does not rely on our annotation framework and is adaptable to other techniques and environments. We believe this study will promote digitalization for higher productivity at industrial sites.

Acknowledgments

This work is partially supported by JSPS KAKENHI (JP22K18422), NEDO (JPNP23003), NICT (222C01, 22609), CSTI SIP3 (JPJ012495), and JST BOOST (JPMJBS2422). We also thank TRUSCO Nakayama Corporation for providing the experimental environment.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

[2] Xinlei Chen and Kaiming He. 2021. Exploring Simple Siamese Representation Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15745–15753. https://doi.org/10.1109/CVPR46437.2021.01549

[3] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M. Alvarez. 2021. Active Learning for Deep Object Detection via Probabilistic Modeling. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 10244–10253. https://doi.org/10.1109/ICCV48922.2021.01010

[4] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. 2018. Modeling Visual Context Is Key to Augmenting Object Detection Datasets. In *Computer Vision - ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*. 375–391. https://doi.org/10.1007/978-3-030-01258-8_23

[5] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. 2017. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1310–1319. https://doi.org/10.1109/ICCV.2017.146

[6] Nathan Elangovan, Ricardo V. Godoy, Felipe Sanches, Ke Wang, Tom White, Patrick Jarvis, and Minas Liarokapis. 2023. On Human Grasping and Manipulation in Kitchens: Automated Annotation, Insights, and Metrics for Effective Data Collection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 11329–11335. https://doi.org/10.1109/ICRA48891.2023.10161171

[7] Georgios Georgakis, Arsalan Mousavian, Alexander Berg, and Jana Kosecka. 2017. Synthesizing Training Data for Object Detection in Indoor Scenes. In *Proceedings of Robotics: Science and Systems*. https://doi.org/10.15607/RSS.2017.XIII.043

[8] Keisuke Higashiura, Kodai Yokoyama, Yusuke Asai, Hironori Shimosato, Kazuma Kano, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2024. Semi-Automated Framework for Digitalizing Multi-Product Warehouses with Large Scale Camera Arrays. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 98–105. https://doi.org/10.1109/PerCom59722.2024.10494498

[9] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. *Ultralytics YOLOv8*. https://github.com/ultralytics/ultralytics

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*. 21–37.

[11] Jiahao Lu, Chong Yin, Oswin Krause, Kenny Erleben, Michael Bachmann Nielsen, and Sune Darkner. 2022. Reducing Annotation Need in Self-explanatory Models for Lung Nodule Diagnosis. In *Interpretability of Machine Intelligence in Medical Image Computing: 5th International Workshop, IMIMIC 2022, Held in Conjunction with MICCAI 2022, Singapore, Singapore, September 22, 2022, Proceedings*. 33–43. https://doi.org/10.1007/978-3-031-17976-1_4

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

[13] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2018. Semantic Foggy Scene Understanding with Synthetic Data. *Int. J. Comput. Vision* 126, 9 (sep 2018), 973–992. https://doi.org/10.1007/s11263-018-1072-8

[14] Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*. 402–419. https://doi.org/10.1007/978-3-030-58536-5_24

[15] Vladyslav Usenko, Nikolaus Demmel, and Daniel Cremers. 2018. The Double Sphere Camera Model. In *2018 International Conference on 3D Vision (3DV)*. 552–560. https://doi.org/10.1109/3DV.2018.00069

[16] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. 2017. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III*. 399–407. https://doi.org/10.1007/978-3-319-66179-7_46

[17] Kodai Yokoyama, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2023. Digitization and Analysis Framework for Warehouse Truck Berth. In *2023 Fourteenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*. 1–4. https://doi.org/10.23919/ICMU58504.2023.10412228

[18] Donggeun Yoo and In So Kweon. 2019. Learning Loss for Active Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 93–102. https://doi.org/10.1109/CVPR.2019.00018