2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Efficient Load Interference Detection with Limited Labeled Data

Shinichi Mae[1] and Hirokatsu Kataoka[2]

***Abstract*— The logistics industry is facing major labor shortages owing to the increasing production volume driven by the continued expansion of e-commerce. This situation has accelerated the development of solutions such as autonomous forklifts. For these forklifts to perform stable material handling, the accurate detection of the load state is essential. However, logistics data often contain sensitive information related to customer products, hindering the collection of comprehensive datasets for developing detection technologies based on machine learning. We propose a method for accurately detecting the position and shape of a load as a mask using limited load data and subsequently identifying the interference state between loads based on the predicted masks. The proposed method leverages instance segmentation pre-trained with formula-driven supervised learning (FDSL) to achieve highly accurate mask prediction, even with limited labeled data for fine-tuning. Pre-training using FDSL leads to a high detection accuracy with a mean average precision (intersection-over-union threshold of 90) of 91.0% using only 400 images. Furthermore, interference detection based on the predicted masks reaches high rates, with a precision of 95.0% and recall of 95.0% on an evaluation set that includes loads with and without interference. Our findings indicate that accurate load interference detection can be achieved with limited labeled data, possibly contributing to the advancement of automation in the logistics industry.**

I. INTRODUCTION

With the rapid expansion of e-commerce, the resulting surge in logistics volume has led to a critical shortage of workers in logistics sites [1]. This pressing issue underscores the immediate need to reduce labor dependence through the automation of logistics operations. In particular, forklifts handle cargo loaded on pallets (flat platforms for transporting goods) and are essential for loading and unloading trucks or shelves in logistics sites. Hence, automating forklifts is expected to drastically reduce labor requirements. However, existing autonomous forklifts require a structured operating environment to ensure stable load handling. Consequently, the deployment of autonomous forklifts to diverse sites is limited, mainly because of the interference between loads, as illustrated in Fig. 1. This interference occurs when vibrations during transport by trucks or other means cause cargo or pallets to shift, resulting in overlapping adjacent loads. Handling interfering loads is challenging with conventional load handling operations, and technologies to detect interference between loads must be devised to promote the adoption of autonomous forklifts.

*This work was not supported by any organization.

[1]TICO-AIST Cooperative Research Laboratory for Advanced Logistics (ALlab), National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan

[2]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki, Japan

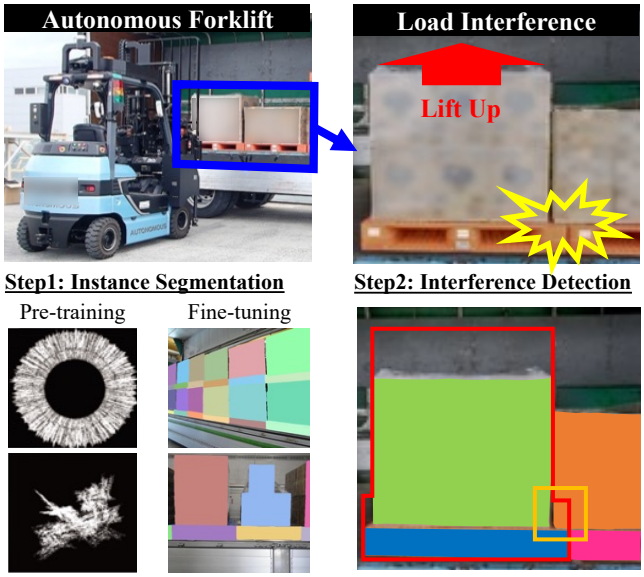

Fig. 1: Detection of load interference in logistics sites. Interference between adjacent loads can pose challenges for material handling. The proposed approach detects load interference in two steps. First, formula-driven supervised learning is used for pre-training, substantially reducing the amount of required load data to build an instance segmentation model that accurately predicts masks for both cargo and pallets. Second, load interference is detected from the predicted masks for cargo and pallets.

To detect load interference, the position and shape of load should be accurately estimated. To this end, an effective method is instance segmentation, which predicts individual regions of target objects in an image as pixel-level masks; thus, it is used in various fields such as autonomous driving and robotic picking [2]–[5]. Typically, instance segmentation models require the collection of images and pixel-level labeling of each object in the images, making dataset construction costly. In logistics, load data often contain sensitive customer product information, which is rarely disclosed owing to confidentiality concerns and further hinders data collection. Alternatively, we apply mosaic processing to sections of load images taken at customer sites and containing product information.

Pre-training on large datasets such as ImageNet allows to compensate for scarce training data. In practice, deep learning models generally show improved performance by pre-training on large datasets and then fine-tuning to the target task. However, many large datasets, including ImageNet, prohibit commercial use owing to personal information protection and fairness assurance. Therefore, careful consideration is needed for using such datasets to implement

pre-training models in products. Kataoka et al. proposed formula-driven supervised learning (FDSL) that uses images generated from mathematical formulas for pre-training to address the above mentioned problems of large datasets [6]–[8]. FDSL can automatically generate images and corresponding labels based on formula parameters, eliminating the labor required for dataset construction. In addition, the generated images are copyright free for commercial use. As FDSL can generate a diverse range of data depending on the formula parameters, it is likely suitable and effective for logistics dataset construction.

We propose the method illustrated in Fig. 1 for detecting load interference by accurately predicting the masks of loads and pallets using FDSL for pre-training. The proposed method can suitably handle the data scarcity in logistics.

The contributions of this study are summarized as follows:

- We constructed an accurate instance segmentation model using a small load dataset and FDSL for pre-training. ExFractalDB-21k (FDSL dataset) was used for pre-training, and only 400 load images were required for fine-tuning, achieving a high prediction accuracy, with 91.0% mean average precision (mAP) at an intersection-over-union threshold of 90.
- For load interference detection, using the masks predicted by the model trained on 400 load images, we achieved a high detection rate with a precision of 95.0% and recall of 95.0% on a test set containing images of loads with and without interference.

II. RELATED WORK

A. Instance segmentation for real-world applications

Instance segmentation involves identifying individual objects within an image and estimating their boundaries at the pixel level. This task plays a critical role in various industrial applications, such as detecting pedestrians and vehicles in autonomous driving or identifying graspable regions of objects in robotic manipulation [4], [5]. However, achieving this requires detailed pixel-level annotations for each object region, making the construction of datasets extremely costly [9]. This challenge is further amplified in domain-specific datasets, such as those for medical or logistics applications, where images must be collected on-site. These images often contain sensitive information, limiting their availability for public release and limiting the scalability of such datasets. Additionally, for real-world applications, it is crucial to develop models capable of accurately segmenting object regions. However, when access to large-scale datasets is limited, building models that can achieve high-precision mask predictions from small datasets becomes essential. Addressing this issue represents a significant challenge in instance segmentation.

B. Pre-training on Large Datasets

When dataset construction is expensive and scarce data are available, like in instance segmentation, pre-training is a convenient approach. In deep learning, pre-training involves training a model on a large dataset (e.g., ImageNet) to perform image classification and then learn the target task. This process helps extract general features from images, thereby enhancing the performance of the target task. Pre-training is effective even when it differs from the target task for fine-tuning. Recently, vision transformers, which are based on the transformer architecture, have outperformed traditional models based on convolutional neural networks [10]. Vision transformers achieve high accuracy by pre-training on large datasets. By using datasets larger than ImageNet, such as JFT-300M/3B [11] and Instagram-3.5B [12], better results have been obtained. However, publicly available datasets like ImageNet are restricted to academic use owing to privacy and fairness concerns. Therefore, using such datasets for commercial purposes requires careful consideration.

C. Synthetic Pre-training for Instance Segmentation

Large-scale datasets commonly used in academic research often come with restrictions on commercial use, posing significant challenges for industrial applications. To address these limitations, synthetic data approaches have been proposed, particularly for segmentation tasks [13], [14]. Typically, these approaches generate synthetic data by placing 3D models and cameras within photorealistic simulation environments. Segmentation masks are then automatically created based on the generated images and the corresponding 3D model shapes. This method efficiently generates diverse, high-precision segmentation data by varying 3D model types and camera placements. However, when applied to a different target domain, the generated data is often unsuitable without adaptation, necessitating the creation of new simulation environments for each domain. This requirement significantly increases the cost of data generation, posing a major challenge. To overcome these challenges, Kataoka et al. introduced Formula-Driven Supervised Learning (FDSL) [7], [8], which leverages images generated from mathematical formulas for pre-training. FDSL circumvents copyright issues commonly associated with large datasets, making it well-suited for commercial use. Additionally, since both images and corresponding labels are automatically generated based on formula parameters, dataset construction is highly cost-efficient. FDSL also enables the creation of diverse datasets by modifying generation parameters, ensuring flexibility across various tasks. Models pre-trained using FDSL have demonstrated accuracy comparable to, or even surpassing, models pre-trained on traditional datasets such as ImageNet in tasks like image classification. However, its applicability to industrial datasets, particularly in specialized domains such as logistics, remains unexplored. In this study, we evaluate the effectiveness of FDSL as a pre-training method for industrial datasets, focusing on load segmentation in logistics environments.

III. METHOD

In this section, we first explain the method for image collection and labeling used in the logistics load dataset for model fine-tuning. Next, we describe model training with

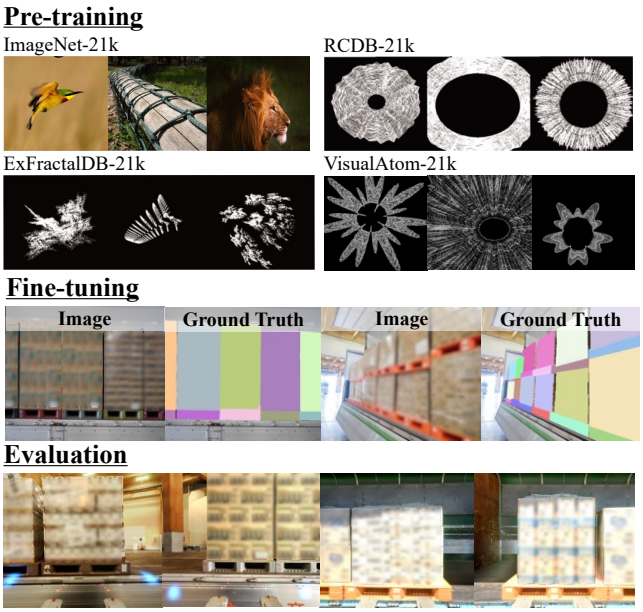

Fig. 2: Samples from pre-training and fine-tuning datasets. For pre-training, we used three types of FDSL datasets and ImageNet-21k as the ground-truth dataset for comparison. For fine-tuning, we used a palletized load segmentation dataset, which consists of load images with annotated regions for cargo and pallets. To evaluate the model, we employed an evaluation set containing images of loads with and without interference.

FDSL for pre-training. Finally, we detail the method for geometrically detecting load interference using the masks of cargos and pallets predicted by the model.

A. Dataset Collection and Labeling

We manually annotated 800 load images with high precision to create a Palletized Load Segmentation dataset (PLSeg). Samples from PLSeg are shown in Fig. 2. The images in PLSeg were captured using digital cameras and cameras mounted on the front of forklifts in both real and simulated logistics sites. The images were annotated using polygon mask segmentation, where the front regions of the cargo and pallets served as ground truths. To detect load interference, accurate mask prediction at the millimeter level was required. Thus, professional annotators carried out this task, and multiple reviewers checked the annotations to ensure high accuracy. For model training, we used small datasets of 200 and 400 images randomly selected from the 800-image dataset to test the segmentation accuracy with scarce data. For evaluation, we used 100 images consisting of 80 and 20 cases without and with interference, respectively. The images were captured by a camera mounted on the front of a forklift. Samples from the evaluation set are also shown in Fig. 2. These test data were used to evaluate load interference detection.

B. Model Training

Here, we provide a detailed explanation of model training. The model backbone was the Swin transformer base [15] based on a transformer architecture, and the head was the

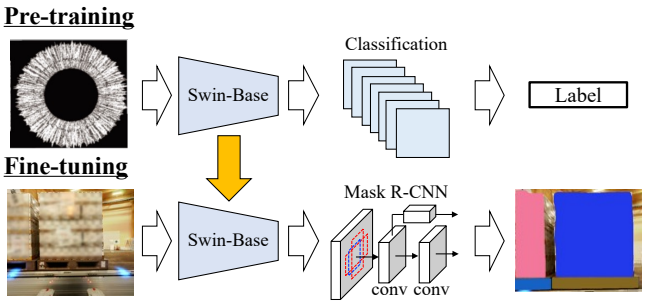

Fig. 3: Model training. The model backbone was first pre-trained on an image classification task using FDSL to learn general features. Then, it was fine-tuned on PLSeg for accurate prediction of masks for loads and pallets.

Mask R-CNN (mask region-based convolutional neural network) [2] commonly used for instance segmentation. FDSL was used for pre-training, and PLSeg was used for fine-tuning. Fig. 3 shows an overview of model training.

1) Pre-training: We used three types of synthetic datasets for pre-training using FDSL: ExFractal Data Base 21k (EFDB-21k) [7], Radial Contour Data Base 21k (RCDB-21k) [7], and Visual Atoms Data Base 21k (VADB-21k) [7]. EFDB-21k comprises synthetic images rendered from 3D fractals from multiple viewpoints. Hence, the model can acquire the ability to broadly recognize shapes and structures of objects, enhancing the extraction of visual features. RCDB-21k comprises images of contour shapes generated by overlaying multiple polygons, enabling specialized feature extraction of contours and shapes of objects. VADB-21k comprises diverse contours and shapes generated from figures created by overlaying sine waves on elliptical orbits. It allows the model to accurately recognize smooth curves and complex shapes. Each FDSL dataset comprises 21,000 classes with 1,000 instances per class. For comparison and validation, we used ImageNet-21k, which is widely used for pre-training image recognition models. The pre-training conditions for each dataset followed the protocols outlined in the corresponding references using PyTorch image models [16].

2) Fine-tuning: For fine-tuning, we used PLSeg. The initial network weights were those learned from each pre-training dataset. Fine-tuning was conducted using the AdamW optimizer with a learning rate of 0.001, batch size of 16, and weight decay of 0.05 for 60 epochs. The model was implemented using the MMDetection toolbox in PyTorch [17].

C. Load Interference Detection

In this section, we describe a method for detecting load interference using the masks of cargos and pallets predicted by the trained model. Fig. 4 shows a schematic of the proposed load interference detection method.

1) Step 1. Palletized load configuration: First, we define each mask predicted by the model as a single load, including every pallet and cargo stacked on it. The centroid coordinates, $\mathbf{c}_i = (x_i, y_i)$, of each pallet mask are calculated, and

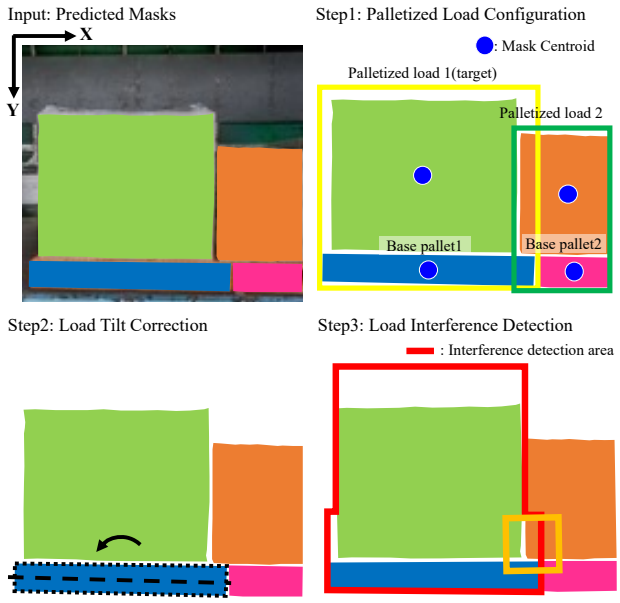

Fig. 4: Load interference detection method. By using the masks of loads and pallets predicted by the instance segmentation model, load interference is detected by correcting the tilt of the load and identifying whether the masks of adjacent loads are included in the area of forklift transit during load lifting.

a search is performed to identify the base pallet as the object whose centroid lies within the following range and has the highest y-axis coordinate value:

$$x_{\min,i} \leq x_j \leq x_{\max,i} \quad \text{and} \quad y_i \leq y_j \leq y_{\max}, \tag{1}$$

where $x_{\min,i}$ and $x_{\max,i}$ are the x-axis boundaries of pallet mask i, and $\mathbf{c}_j = (x_j, y_j)$ are the centroid coordinates of pallet mask j. If no other pallets fall within this range, mask i is designated as the base pallet.

Next, for each base pallet, we identify all other masks corresponding to loads or pallets recognizing them as items stacked on the base pallet. These items have centroids within the following range:

$$x_i - \frac{w_i}{10} \leq x \leq x_i + \frac{w_i}{10} \quad \text{and} \quad y_i \leq y \leq y_{\max}. \tag{2}$$

Finally, the palletized load whose base pallet has the centroid with the x-axis coordinate closest to the image center is designated as the target load. This criterion is applicable because the currently handled load appears near the center in images captured by a forward-facing forklift camera.

2) Step 2. Load tilt Correction: Next, we correct the load tilt. Pallets used in logistics are rectangular, with a longer horizontal side when viewed from the front. To confirm this constraint, the vertex coordinates of the base pallet mask of the target load detected in step 1 are used for principal component analysis. The direction of the first principal component is estimated to align with the long side of the pallet, obtaining the angle needed to align this direction with the x axis. Based on this angle, all the cargo and pallet masks are rotated to correct the load tilt.

TABLE I: Results for pre-training and fine-tuning for instance segmentation with limited labeled load data.

Pre-training		Fine-tuning		mAP_{75}	mAP_{90}
Dataset	Data type	Dataset	Samples		
Scratch	–			76.6	46.8
ImageNet-21k	Real			**92.5**	**80.3**
EFDB-21k	FDSL	PLSeg	200	90.2	78.9
RCDB-21k	FDSL			89.3	76.4
VADB-21k	FDSL			88.3	64.5
Scratch	–			94.1	71.1
ImageNet-21k	Real			98.7	**91.7**
EFDB-21k	FDSL	PLSeg	400	**99.2**	91.0
RCDB-21k	FDSL			98.7	86.3
VADB-21k	FDSL			**99.2**	84.7
Scratch	–			98.6	85.1
ImageNet-21k	Real			**99.4**	**98.7**
EFDB-21k	FDSL	PLSeg	800	99.1	94.1
RCDB-21k	FDSL			99.3	90.9
VADB-21k	FDSL			98.3	90.7

3) Step 3. Load interference detection: Finally, we define the interference detection area as the transit region of the target load when lifted by a forklift. As shown in the image, when the load is lifted by the forklift, it moves along the negative y-axis direction. Therefore, the interference detection area is defined as the region through which the mask of the target load passes when shifted by a certain number of pixels along the negative y-axis direction. In this study, we set the displacement to correspond to 100 mm, which is a typical value for a forklift lift. Interference is detected if the mask of another load is present within the defined interference detection area.

IV. EXPERIMENTS AND RESULTS

We examined whether high-accuracy mask prediction of cargos and pallets could be achieved with scarce labeled load data by using FDSL for pre-training. Then, we verified whether the predicted masks could be used to accurately detect load interference.

A. Prediction of Palletized Load Mask with Limited Data

The predicted masks were evaluated in terms of mAP, a common metric for instance segmentation. Given the required highly accurate mask prediction, we evaluated high intersection-over-union thresholds of 75 and 90 for mAP (mAP_{75} and mAP_{90}, respectively). Table I lists the prediction accuracy for each evaluated pre-training and fine-tuning method. First, when fine-tuning was performed with 800 load images, the highest mAP_{75} and mAP_{90} scores were achieved on ImageNet-21k. Next, when fine-tuning was performed with 400 load images, the highest mAP_{75} scores were achieved on EFDB-21k and VADB-21k. In terms of mAP_{90}, EFDB-21k led to scores comparable to those obtained from ImageNet-21k. Lastly, when fine-tuning was performed with 200 load images, the highest mAP_{75} and mAP_{90} scores were achieved on ImageNet-21k, but none of the pre-training datasets led to a mAP_{90} value above 90%. Overall, while ImageNet-21k, a real image dataset, is

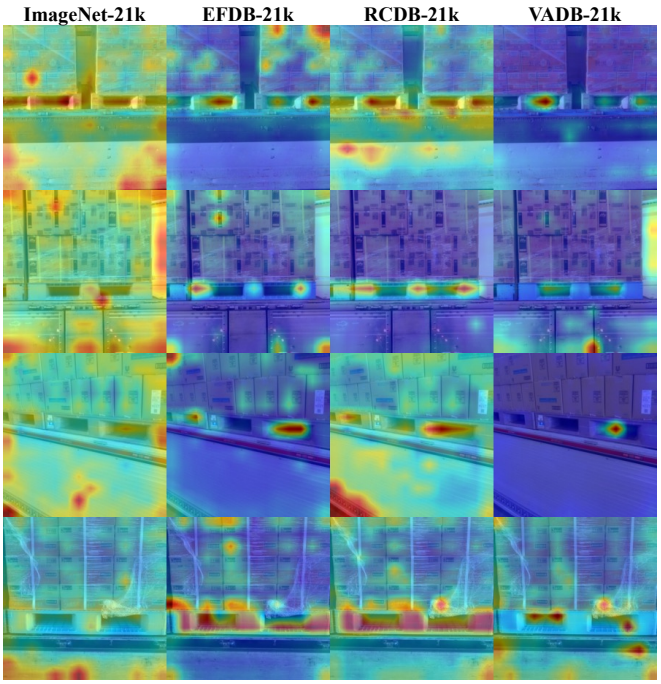

Fig. 5: Attention maps for load images obtained from vision transformers pre-trained on ImageNet-21k, EFDB-21k, RCDB-21k, and VADB-21k. The model pre-trained on ImageNet-21k broadly distributes its attention across the entire image, whereas the models pre-trained on the three FDSL datasets tend to focus on the pallet area. Notably, the model pre-trained on EFDB-21k shows strong attention to the edges of pallets and cargo.

effective when there are sufficient data, FDSL is effective even for 400 images. Among the FDSL datasets used in this study, EFDB-21k showed the best results. Next, the pre-training effect of FDSL was analyzed using attention maps, which visually indicate the parts of an input image that a model focuses on. For analysis, we pre-trained vision transformers on the ImageNet-21k, EFDB-21k, RCDB-21k, and VADB-21k datasets and then examined the attention maps when each model was fed with load images. This evaluation provided the image areas each model focused on during pre-training. Fig. 5 shows the attention maps of the evaluated models. The model pre-trained on ImageNet-21k showed a broad distribution of attention over the entire image, focusing mainly on the edges. By contrast, the models pre-trained on the three FDSL datasets tended to focus on areas near the pallets. Notably, the model pre-trained on EFDB-21k exhibited a strong focus on areas near the edges of both pallets and loads. These results suggest that pre-training on EFDB-21k enabled the model to extract features focusing on both the pallets and loads captured in the images. As a result, even when trained on few load images, the model's ability to detect loads and pallets was improved.

B. Load Interference Detection

Next, we used the masks predicted by each model constructed in section IV-A to verify the detection of load interference. For evaluation, we considered data of loads with and without interference to assess the corresponding detection rate in terms of precision and recall. Table II lists the interference detection rates for each pre-trained model, and Fig. 6 shows the detection results on input images. First, when fine-tuning was performed using 800 load images, the model pre-trained on EFDB-21k achieved a precision of 0.95 and recall of 1.00, confirming that interference could be accurately detected using the predicted masks for loads and pallets. Even when fine-tuning was performed using 400 load images, the model pre-trained on EFDB-21k achieved a precision of 0.95 and recall of 0.95, demonstrating that interference could be detected with limited labeled data. By contrast, when fine-tuning was performed with 200 load images, all the pre-training datasets led to low detection rates, with precision and recall values below 0.8, indicating that these models were unsuitable for load interference detection. While the evaluation of load and pallet mask prediction showed that the model pre-trained on ImageNet-21k performed better than that pre-trained on EFDB-21k, the latter dataset provided better interference detection, achieving higher detection rates. This result suggests that the mask prediction accuracy between adjacent loads is crucial for interference detection. As shown in Fig. 6, the model pre-trained on EFDB-21k exhibited higher mask accuracy between adjacent loads, which likely contributed to the improved interference detection. As discussed in section IV-A, the model pre-trained on EFDB-21k tended to focus attention near the edges of loads and pallets, possibly enhancing the mask prediction accuracy near adjacent loads and thereby increasing the interference detection rate.

TABLE II: Results for pre-training and fine-tuning for load interference detection with limited labeled load data.

Pre-training		Fine-tuning		Precision	Recall
Dataset	Data type	Dataset	Samples		
Scratch	–			0.00	0.00
ImageNet-21k	Real			0.73	**0.55**
EFDB-21k	FDSL	PLSeg	200	**0.80**	0.40
RCDB-21k	FDSL			0.14	0.50
VADB-21k	FDSL			0.25	0.17
Scratch	–			0.74	0.85
ImageNet-21k	Real			0.86	**0.95**
EFDB-21k	FDSL	PLSeg	400	**0.95**	**0.95**
RCDB-21k	FDSL			0.89	0.80
VADB-21k	FDSL			0.86	0.60
Scratch	–			0.88	0.75
ImageNet-21k	Real			0.83	**1.00**
EFDB-21k	FDSL	PLSeg	800	**0.95**	**1.00**
RCDB-21k	FDSL			0.83	**1.00**
VADB-21k	FDSL			0.89	0.85

V. CONCLUSIONS

We aimed to promote the adoption of autonomous forklifts by focusing on the accurate detection of load interference in logistics sites. We achieved high-accuracy mask prediction for loads and pallets from scarce training data (400 images) and confirmed that the predicted masks could be used to detect load interference. For pre-training, we employed FDSL, which generates synthetic datasets from mathematical

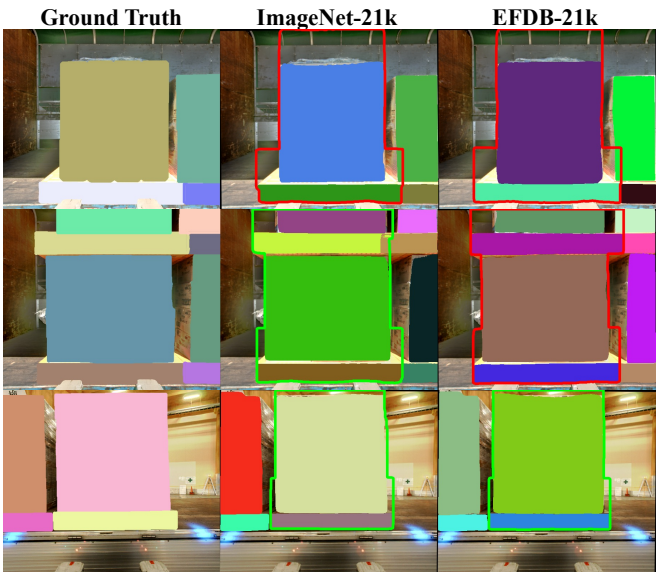

Fig. 6: Results of load interference detection using masks predicted by model pre-trained on ImageNet-21k and EFDB-21k and fine-tuned with 400 load images. The red and green outlines represent the boundaries of the area through which the load passes when lifted, with red indicating detected load interference and green outline indicating no detected interference. The first and second rows show detection results for interfering loads. In the first row, interference is detected in both cases, but in the second row, interference is detected only when using the model pre-trained on EFDB-21k. The third row shows detection results for a load without interference.

formulas, to predict load and pallet masks with high accuracy using limited data. For fine-tuning, we used a dataset collected from logistics sites consisting of palletized load images with precisely annotated cargo and pallet regions. Using the EFDB-21k FDSL dataset for pre-training, we achieved a high detection performance with mAP_{90} of 91.0% using only 400 load images. To detect load interference, we utilized a method that determined whether the mask of a neighboring load was included in the handling region of the forklift when lifting a load. Using masks predicted by the model pre-trained on EFDB-21k and fine-tuning on 400 load images, we achieved a high interference detection rate, with a precision of 0.95 and recall of 0.95, confirming that the proposed method accurately detected load interference. The proposed method can detect load interference occurring in logistics sites, possibly expanding the environments where autonomous forklifts can be deployed. Various logistics systems using pallets may be endowed with stable handling by simply installing cameras. In future work, a load interference detection system using the proposed method will be implemented on autonomous forklifts operating in real-world environments to demonstrate real-time interference detection. As the proposed method occasionally misidentified loads with interference, the underlying model should be further improved. For instance, fine-tuning a foundational model like Segment Anything [18] with load images may enable general detection applicable to logistics sites.

REFERENCES

[1] W. Bank, *Digital Platforms and the Future of E-commerce*. World Bank Publications, 2020.

[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[3] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

[4] Q. Yang, J. Peng, D. Chen, and H. Zhang, "Road scene instance segmentation based on improved solov2," *Electronics*, vol. 12, no. 19, p. 4169, 2023.

[5] J. Dirr, J. C. Bauer, D. Gebauer, and R. Daub, "Cut-paste image generation for instance segmentation for robotic picking of industrial parts," *The International Journal of Advanced Manufacturing Technology*, vol. 130, no. 1, pp. 191–201, 2024.

[6] H. Kataoka, K. Okayasu, A. Matsumoto, E. Yamagata, R. Yamada, N. Inoue, A. Nakamura, and Y. Satoh, "Pre-training without natural images," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[7] H. Kataoka, R. Hayamizu, R. Yamada, K. Nakashima, S. Takashima, X. Zhang, E. J. Martinez-Noriega, N. Inoue, and R. Yokota, "Replacing labeled real-image datasets with auto-generated contours," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21232–21241, 2022.

[8] S. Takashima, R. Hayamizu, N. Inoue, H. Kataoka, and R. Yokota, "Visual atoms: Pre-training vision transformers with sinusoidal waves," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18579–18588, 2023.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[10] 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," *ICLR*, 2021.

[11] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

[12] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018.

[13] Y. Cabon, N. Murray, and M. Humenberger, "Virtual kitti 2," *arXiv preprint arXiv:2001.10773*, 2020.

[14] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing, "Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3105–3115, 2019.

[15] 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," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

[16] R. Wightman, "Pytorch image models," 2019.

[17] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.