Semi-Automated Framework for Digitalizing Multi-Product Warehouses with Large Scale Camera Arrays

Keisuke Higashiura*, Kodai Yokoyama*, Yusuke Asai*, Hironori Shimosato*,
Kazuma Kano*, Shin Katayama*, Kenta Urano*, Takuro Yonezawa*, Nobuo Kawaguchi*†

*Graduate School of Engineering, Nagoya University, Japan
†Institutes of Innovation for Future Society, Nagoya University, Japan
Email : urachan, yokochin, asayu, shimo, kazuma, shinsan, vrano @ucl.nuee.nagoya-u.ac.jp,
takuro, kawaguti@nagoya-u.jp

***Abstract*—As global demand for logistics continues to grow, optimizing the automation and efficiency of distribution warehouse operations is of paramount importance. Digitalizing warehouse environments, which refers to the process of sensing the physical space and extracting meaningful information from the obtained data, offers a promising solution to this challenge. However, converting raw warehouse data, such as video footage captured inside the warehouse, into actionable metadata (e.g., tracking the movement paths of workers and products or analyzing the usage patterns of different warehouse locations) often necessitates significant human intervention. The rise of machine learning further complicates this, as it requires the manual preparation of extensive training datasets. In this paper, we introduce a framework that semi-automates the digitalization process in complex warehouse settings. This framework employs dense optical flow and representation learning to autonomously segment warehouse objects and cluster similar objects, thereby substantially cutting down on annotation costs. To evaluate our approach, we constructed a large-scale data collection platform with over 60 fixed cameras in a real-world logistics warehouse, and the video data from this platform was then applied to our framework. Our evaluations indicate that our method markedly reduces both the time and resources required for warehouse digitalization using the captured video data.**

***Index Terms*—smart warehouse, data digitalization, digital twin**

I. Introduction

The global logistics market expansion necessitates automated and efficient business processes in warehouses. Studies have focused on various methods to enhance warehouse operations efficiency [1]–[3], notably digital twins. This technology replicates real-world objects, processes, and human activities in cyberspace, facilitating efficient warehouse management [4]. Digital twins allow for the optimization of worker allocation, product placement, and transportation routes in warehouses. Unlike physical trials, which are disruptive and time-consuming, digital twins enable cost-effective, non-intrusive simulations in cyberspace.

Building a digital twin that is faithful to the physical space requires an accurate digitalization of the physical space, which refers to the process of sensing the physical space and extracting meaningful data from the acquired data. However, deriving useful metadata from raw data, such as warehouse videos for tracking movement or analyzing space usage, requires extensive human effort. For example, object recognition in warehouses via machine learning demands extensive data preparation and annotation. In particular, logistics warehouses are highly complex environments. Since they contain millions of diverse objects, they require more annotation costs than typical environments. This is a significant challenge in the digital twinning of logistics warehouses.

In this paper, we propose a semi-automated framework for digitalizing the physical space of logistics warehouses with a wide variety of objects. This framework achieves efficient annotation of warehouse objects with minimal human resources by automatically segmenting warehouse objects based on motion detection methods and estimating objects of the same class through representation learning and clustering. In this study, we constructed a large-scale data collection infrastructure composed of more than 60 fixed cameras in an actual logistics warehouse in Aichi Prefecture, Japan, and digitalized the logistics warehouse using warehouse footage. We also applied this framework to the recognition of workers and handpallets performing tasks in a logistics warehouse and showed that this method can reduce the cost of object annotation by 98.6%.

The contributions of this paper are as follows:

- In an actual logistics warehouse, we constructed the largest-scale data collection infrastructure to date, composed of more than 60 fixed cameras, and digitalized the logistics warehouse using warehouse footage.
- Based on our digitalization case study, we present the challenges in digitalizing logistics warehouses and building a digital twin.
- We proposed a method to significantly reduce the annotation costs of warehouse objects, which is a major challenge in the digital twinning of logistics warehouses.

This paper first presents research relevant to the topic

(Section II). Next, an overview of the large-scale camera arrays we have built (III), a pre-study of warehouse digitalization using it (IV), and the challenges in warehouse digitalization identified from this study (V) are presented. The architecture of the proposed framework to solve these challenges is then presented (VI), followed by experiments (VII), evaluations (VIII) and discussions (IX) to demonstrate its effectiveness.

II. Related Work

In recent years, research aimed at improving the efficiency of operations within logistics warehouses includes the use of digital twins for the introduction of Autonomous Mobile Robots (AMR) to enhance warehouse utilization [5], task scheduling and optimization of goods storage warehouses using digital twins [6], [7]. Zhou et al. [8] proposed a framework called SOD-DT for constructing digital twins by extracting small objects present in the warehouse. For this method, annotations were made on 5000 images to train the recognition model. The amount of time they spent on this task is not clear in their paper, but annotating images generally takes a lot of time and requires a significant amount of human resources.

Most deep learning based recognition methods have the weakness of requiring a large amount of training data to train the recognition model. However, generic models capable of zero-shot object recognition [9]–[11] have emerged in the last 1-2 years. These methods are very promising as they do not have to spend any annotation costs for recognizing specific objects, but building a generic model requires huge computational resources, making the construction cost extremely high. For example, one of the latest methods for achieving zero-shot panoptic segmentation called ODISE [10] generates training images by augmenting 1024^2 original images and then trains the model for 90k iterations using 32 NVIDIA V100 GPUs, requiring a massive dataset and computational resources. Additionally, the warehouse environment we are targeting is a complex scene with a wide variety of objects, and the recognition accuracy in such an environment is unknown.

Research has also been conducted to partially automate annotations to reduce the workload, including methods for estimating the malignancy of lung nodule diagnosis with few annotations using self-supervised learning [12], and methods for automatically annotating the state of hands during cooking [13]. However, these methods are dependent on specific domains or datasets and are difficult to apply directly to our environment. Furthermore, an approach called active learning, which prioritizes annotating data that is thought to be highly effective for learning and reduces the number of annotations [14]–[16], has also been proposed. These methods achieve higher accuracy with fewer annotations, but it is stated that annotations of half or several thousand samples of the entire dataset are still necessary. In the warehouse environment we are targeting, with its diverse range of objects, this could potentially be a significant burden.

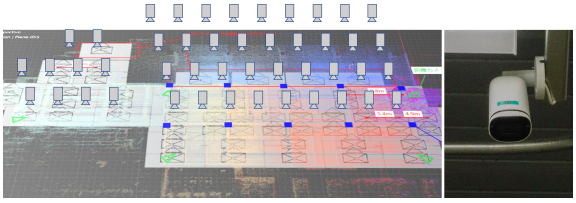

Fig. 1. Schematic of the camera installation locations and installed camera

Fig. 2. Application for lining up images captured by cameras

III. Building a Large Scale Camera Array for Warehouse Digitalization

In this section, we describe building a large-scale camera arrays in a logistics warehouse in Aichi Prefecture, Japan, to create its digital twin. The cameras equipped in the warehouse capture about 1.2TB of daily video, totaling over 200TB in 10 months. Privacy concerns are addressed by informing employees about the cameras and obtaining prior approval for data collection. Shao et al. [17] highlight accuracy as key in digital twin development, advising minimalism to prevent errors. Our camera arrays, covering the entire warehouse, captures detailed, essential data without redundancy. Our approach, which focuses solely on camera-based data acquisition, is ideal for creating accurate digital twins.

Our logistics warehouse is equipped with a camera array of 66 fixed-point cameras, as shown in Fig. 1. These cameras installed at strategic intervals on the ceiling and provide comprehensive coverage with some offering a direct downward view and others a bird's eye view. This setup includes cameras positioned inside the warehouse and at truck berth, which is important place for incoming and outgoing deliveries, for monitoring loading and unloading activities. The camera we used is H.View HV-800G2A5. Recording resolution was set to 1920×1080, and frame rate was set to 5 fps. All cameras are network-connected, with data stored on the warehouse's network storage.

We apply several post-processing techniques to enhance the usability of videos collected for collaboratively sensing a logistics warehouse using multiple cameras. Firstly, accurate synchronization of timestamps across videos recorded by each camera is essential. Each camera periodically synchronizes its time with a Network Time Protocol server. Additionally, timestamps indicating the recording time of each frame are printed in the videos and Optical Character Recognition (OCR) is applied to extract these timestamps. Secondly, the videos often exhibit lens-induced distortion. To correct this and fa-

Fig. 3. Visualization of moving objects' paths

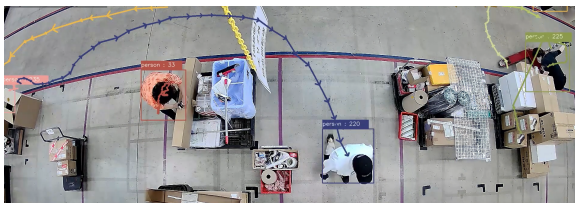

Fig. 4. Example of the object tracking in the warehouse

Fig. 5. Example of the recognition using instance segmentation

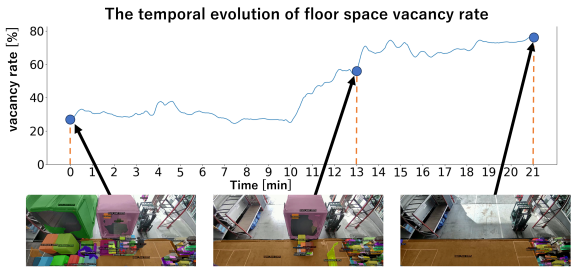

Fig. 6. Temporal evolution of floor vacancy rate in a particular truck berth

cilitate effective video stitching, we use the camera's intrinsic parameters to perform distortion correction. Finally, to provide a comprehensive view of the warehouse, we stitch these corrected videos. This is achieved using a specially developed application shown in Fig. 2, which allows for spatial merging of the footage from each camera onto a 3D model of the warehouse, assigning global coordinates within the warehouse to each video segment.

IV. Pre-Study for Logistics Warehouse Digitalization

In this study, we explore the digitalization of logistics warehouses using our large-scale camera arrays to assess its effectiveness and identify challenges. The study includes two primary analyses.

Firstly, we analyze object movement within the warehouse. By stitching together video footage, we digitally captured object movements and analyzed their paths by computing differences between consecutive frames as shown in Fig. 3. This analysis provides insights for optimizing transport routes and strategic placement. Furthermore, we developed an object tracking system employing YOLOv8 [18] and OCSORT [19] for detection and tracking. An example of the tracking result is shown in Fig. 4. This system can be extended for warehouse-wide tracking, integrating stitching processes described in Section III.

Secondly, we conduct truck berth analysis to understand truck berth utilization. By analyzing footage from truck berth cameras, we assess floor conditions using instance segmentation, which identifies and classifies objects and generates segmentation masks as shown in Fig. 5. We classified 14 frequent warehouse classes using Mask R-CNN [20] trained on 1791 annotated images, resulting in a total of 179,324 patterns across all classes. This analysis helps in estimating the utilization of truck berth floor space by counting pixels recognized as floor area. Fig. 6 shows the temporal evolution of the floor space vacancy rate that reveals the condition and efficiency of truck berth operations.

V. Challenges of Logistics Warehouse Digitalization

In our study of logistics warehouse digitalization detailed in Section IV, and through the analysis of videos over 10 months, we identified key challenges essential for precise digitalization and digital twin construction.

A significant challenge is dealing with the wide variety of products in warehouses. These products vary greatly in size, shape, and material, requiring extensive annotation for accurate detection. In the truck berth analysis described in Section IV, around 180,000 annotations were manually generated, which still did not achieve practical accuracy. Generating sufficient annotations for all product patterns is crucial but demands substantial human resources.

Another challenge is the dynamic nature of the warehouse environment. Factors like seasonal changes in products, and alterations in warehouse layout to enhance operational efficiency, mean that a digital twin based on data from a specific time may become outdated as the environment evolves. Therefore, continuous digitalization of the physical space and its regular integration into the digital twin is necessary. This also implies a need for ongoing annotation, especially when encountering unknown objects.

VI. Framework for Semi-Automated Warehouse Digitalization

As described in Section V, the problem of requiring a large amount of annotation consumes a lot of human resources and significantly increases the cost of digitalization. In this section, we propose a semi-automated annotation framework for the digitalization of the warehouse to overcome the challenges.

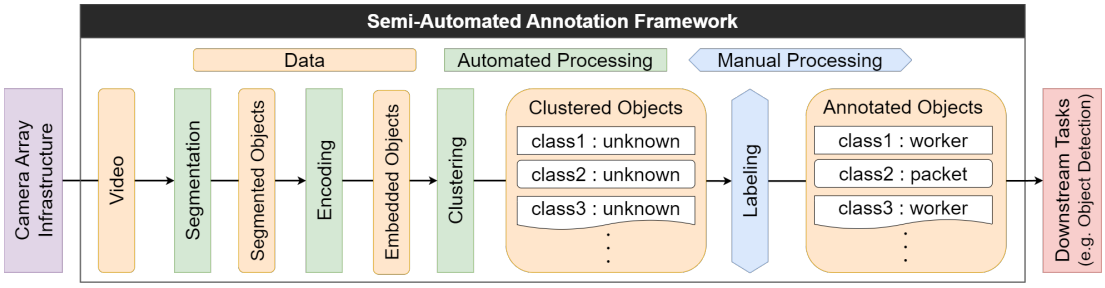

Fig. 7. Overview of the proposed framework

Fig. 8. Example of the segmentation using optical flow

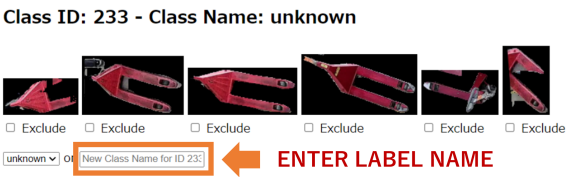

Fig. 9. Screenshot of the labeling tool

Fig. 7 shows the overview of the proposed framework. This framework aims to reduce the annotation cost of warehouse objects, enabling the digitalization of the warehouse with less effort. It consists of four main steps: segmentation, encoding, clustering, and labeling.

A. Segmentation

In this step, we perform segmentation of warehouse objects, which is an essential part of the annotation process. We conduct object segmentation by applying a moving object detection method. many methods have been proposed for detecting moving objects in videos and segmenting them, including methods based on frame subtraction and optical flow and so on. Fig. 8 shows an example of segmenting moving objects in a video using optical flow. The method we use in this study is discussed in Section VII-A1.

B. Encoding

In this step, segmented objects are encoded into embedded representations. The purpose of this step is to map similar objects to similar embedded representations and form clusters of similar objects in the feature space. This concept, known as metric learning or distance metric learning, is widely used in tasks like face recognition and image searching. In this paper, we implement and compare two types of encoders: one based on deep distance learning and the other based on autoencoder. Autoencoder is a method for dimensionality reduction by extracting essential features from the input, and they have the effect of mapping similar objects to close positions in the feature space.

C. Clustering

In this step, we apply clustering to the embedded representations of the objects and group objects that are estimated to belong to the same class. The goal of this step is to group objects that are estimated to be of the same class and to reduce the amount of work required for labeling. Before clustering, we apply a dimensionality reduction method to the embedded representations obtained from the encoder to extract more essential features and reduce computational costs. In this paper, we implement and compare two representative dimensionality reduction methods: UMAP [21] and PCA. Additionally, we implement and compare two clustering methods: DBSCAN [22] and K-Means.

D. Labeling

In this step, we label the clustered objects to complete the annotation. While all the previous steps are automated, this step requires manual work. For labeling, we implemented a labeling tool that works in a web browser. In this tool, objects are displayed in groups obtained in the clustering step, and the user instructs which label to assign to them. Some objects might be misclassified due to errors in class estimation. For these cases, a checkbox to exclude the object is provided below each object image and annotators can exclude that object from the group by checking this box. While this feature may increase the amount of labeling work, it contributes to improving labeling accuracy. Fig. 9 shows a screenshot of this labeling tool.

VII. Experiment

In this section, we describe experiments conducted to verify the effectiveness of the proposed framework. For these experiments, we applied the proposed framework to data collected from our large-scale camera arrays to verify the effect of reducing annotation costs. For this experiment, we focus on annotations of workers and handpallets, the most frequently appearing objects in the warehouse.

A. Implementation of Each Component in the Framework

1) Segmentation: Before the experiments, we considered several methods for segmentation. This time, we applied the

frame subtraction method, Farneback method [23], and RAFT [24] to actual footage taken inside a warehouse and checked the accuracy of the segmentation. The Farneback method is a conventional approach for calculating optical flow, and RAFT is an optical flow estimation method based on deep learning. Using the frame subtraction method, we observed significant effects of brightness changes over time, resulting in a large amount of noise data. The Farneback method exhibited some robustness to brightness changes and produced less noise data, but the contours of the segmentation masks were not accurate. RAFT, being a deep learning-based method, incurred high computational costs but had the best robustness to brightness changes and segmentation accuracy among these methods. Since segmentation accuracy directly affects annotation quality, it is considered important to prioritize accuracy despite the higher computational cost. For these reasons, in this paper, we adopted RAFT as the segmentation method. With RAFT-based segmentation, we used a pre-trained model to estimate optical flow for each frame, and then generated segmentation masks for the moving parts identified as pixels where the output flow magnitude was positive. The pre-trained model used was officially provided and it was trained using the FlyingThings [25], which is a large-scale dataset to enable training and evaluating scene flow methods. Furthermore, the contours of the generated mask image are extracted, and based on this contour information, the object mask is separated for each instance. At this time, masks with pixel counts less than 0.2% of the entire image are removed as noise. Also, during the evaluation of segmentation methods, it was found that RAFT mistakenly detects the black margins in the frame caused by distortion correction as moving objects, so objects with an average brightness value of less than 10 were also removed as noise.

2) Encoding: In this experiment, we prepared two types of encoders for comparison: SimSiam [26], a deep distance learning method, and Vision Transformer based Autoencoder(ViTAE). ViTAE is an autoencoder that includes the Transformer encoder of Vision Transformer [27] as part of its encoder, and the encoder of ViTAE consists of an MLP head composed of a fully connected layer and a GeLU layer, connected ahead of the Transformer encoder. Furthermore, the decoder is composed of fully connected layers, deconvolution layers, and GeLU layers, expanding the dimension of the embedded representation to obtain an output of the same dimension as the input. During training, the goal is to minimize the reconstruction loss of the input and output. The implementation and hyperparameters of SimSiam were based on the official version, with the only modification being a batch size change to 128. ViTAE was implemented using PyTorch, with the dimension of the embedding representation set to 2048, the default value in SimSiam. The optimizer used was Adam, the learning rate was set to the PyTorch default of 0.001, and the batch size was also set to 128. As explained in Section VII-A1, we used RAFT to extract 72465 objects from videos to create training data and conducted training for 100 epochs each. However, the pre-trained parameters officially provided were applied to the Transformer encoder part, and the weights were fixed during training.

3) Clustering: In this experiment, we implemented UMAP and PCA as dimentionality reduction methods before clustering, and conducted a comparative study. This time, the 2048-dimensional embedded representations obtained from the encoder described in Section VII-A2 were compressed to 512 dimensions. UMAP has typical parameters such as $\mathrm{n\_neighbors}$ and $\mathrm{min\_dist}$. $\mathrm{n\_neighbors}$ is a parameter that determines how much emphasis is placed on the local and global structures of the data in dimension reduction. $\mathrm{min\_dist}$ represents the minimum possible distance between data points and functions similarly to $\mathrm{n\_neighbor}$. The bigger the values of both, the more the dimentionality reduction emphasizes the global structure. In this study, we set $\mathrm{n\_neighbors} = 10$ and $\mathrm{min\_dist} = 0$ for the experiments. Additionally, we implemented DBSCAN and K-Means as clustering methods, and conducted a comparative study. Typical parameters of DBSCAN include eps and $\mathrm{min\_samples}$. eps represents the maximum distance at which one data point is considered to be in the neighborhood of another during cluster formation, and is the most critical parameter in DBSCAN. $\mathrm{min\_sample}$ is the minimum number of data points required around a certain data point to form a cluster. This number includes the data point itself, and those not meeting this criterion are treated as noise. For this experiment, we set $\mathrm{eps} = 0.1$ when the encoder was SimSiam, $\mathrm{eps} = 0.2$ when it was ViTAE, and $\mathrm{min\_sample} = 0$ for both. Furthermore, the number of clusters in K-Means was set to 200 when using SimSiam as the encoder, and 400 for ViTAE. These clustering components were implemented using scikit-learn.

B. Annotation Using the Proposed Framework

We applied the proposed framework and performed annotation for the workers and handpallets. In this experiment, videos collected from 21 cameras included in the large-scale camera arrays were used as input. Each component processed the input images, generating 4578 annotations for workers and 1309 annotations for handpallets. The annotations were performed by a single annotator. Additionally, as explained in SectionVI-D, we used checkboxes to remove misclassified objects from the cluster, and instructed the annotator to remove objects that were clearly misclassified or of obviously poor quality. However, checking each object meticulously would require a considerable amount of time, so we also instructed to ignore objects unless they were clearly erroneous. Additionally, in order to obtain a metric for annotation cost when using the proposed framework, the time required for labeling was measured.

C. Training the Recognition Model Using Generated Annotations

Using the annotations generated in Section VII-B, we trained a recognition model for the workers and handpallets. For this experiment, we used Faster R-CNN [28] implemented in Detectron2 [29] as the recognition model, which is a

deep learning based framework for object detection tasks, estimating the class and bounding box of instances. In this experiment, to train the recognition model, we performed training ranging from 5000 to 40000 iterations. After training the model, we applied it to dataset for evaluation to evaluate the model's accuracy. Details on model evaluation are discussed in Section VIII.

D. Preparation of Baseline Method

As a baseline method against our proposed framework, we created a recognition model using the typical fully manual annotation method. For frames of videos collected with the large-scale camera arrays, workers and handpallets were manually annotated, and the same number of annotations as created by the proposed framework was generated. The manual annotation was conducted by two individuals regularly involved in such work. The annotators marked the workers and the handpallets in the images by placing bounding boxes around them. Under the same conditions as in Section VII-C, a recognition model for the workers and the handpallets was created. After training the model, it was applied to unknown data, and a comparison was made with the model created using the proposed framework.

VIII. Evaluation

A. Evaluation Metrics

We employ two primary metrics to assess the proposed method. Firstly, we utilize Average Precision (AP) to quantitatively evaluate recognition task performance. AP, averaging accuracy and recall at various thresholds, provides a single value representing the model's recognition performance. High AP values indicate high object recognition accuracy and good quality of the dataset used for training. Secondly, we assess the efficiency of model creation by considering the amount of time annotators spent intervening. The capability to develop a high-quality model within a short timeframe is crucial for real-world applications. By evaluating these metrics, we comprehensively assess the performance and efficiency of the proposed method.

B. Evaluation Results

For evaluation, annotations were manually performed in the same way as the baseline method, generating 863 worker and 256 handpallet annotations. The results of evaluating the baseline and proposed methods are shown in Table I. For combinations where results are not reported, the clustering resulted in either all objects being concentrated in a single cluster, or in a distribution where each cluster contained only 1 to 2 objects, making it impossible to perform labeling. AP50 and AP75 mean that the Intersection over Union (IoU) threshold used to calculate AP is 50% and 75%, respectively, and the higher these numbers, the stricter the evaluation metric. The AP in the table is the average value after calculating the AP by changing the IoU threshold from 0.50 to 0.95 in increments of 0.05. The time required in the table indicates the time it took to create the training data ($= 4578 + 1309 = 5887$ annotations) for each method.

IX. Discussion

A. Comparison between Baseline and Proposed Method

The recognition accuracy of the worker and handpallet in all evaluation metrics was somewhat superior in the baseline method compared to the proposed method, with a difference of up to 17.5pt at maximum. However, in some evaluation metrics, values close to the baseline were obtained; for example, in the combination of RAFT, SimSiam, UMAP, and K-Means of the proposed method, the AP50 of the worker was about 4pt of the baseline. The main reason for the proposed method not reaching the accuracy of the baseline is attributed to errors in optical flow estimation and noise inclusion in the generated clusters, resulting in some incorrect annotation results being included in the dataset, thus degrading the quality of the training data. Regarding the time taken for annotation, the baseline method took 626 minutes, whereas the proposed method was completed within a maximum of 24 minutes, demonstrating the effectiveness of the proposed method. In every combination of results obtained, the time required was reduced by over 96% compared to the baseline, and the highest-performing combination of RAFT, SimSiam, UMAP, and K-Means achieved a reduction of about 98.6%. In addition, in the present task, objects other than those that were clearly misclassified or of poor quality were not excluded and were labelled as they were. If the annotator performed the exclusion task more carefully, the quality of the dataset could be improved in exchange for the increased work time. There exists a trade-off between working time and dataset quality, necessitating adjustment according to the use case.

B. Comparison of Different Combinations of Method in each Component

In this experiment, we found that RAFT, SimSiam, UMAP, and K-Means generally yielded the highest accuracy metrics. Conversely, when DBSCAN was used in three out of four configurations, it either led to the formation of a single dominant cluster or distributed objects evenly across multiple clusters, with each cluster containing only 1-2 items. This resulted in ineffective groupings. DBSCAN's density-based approach often groups loosely related objects into a single cluster if these clusters are dense and widespread, which likely contributed to this issue. Furthermore, UMAP effectively maintains both local and global data structures during dimension reduction, preserving the overall structure post-reduction. On the other hand, PCA focuses on maximizing data variance for dimension reduction, which may have led to the collapse of feature distribution when used in combination with certain encoders. Interestingly, configurations using DBSCAN with ViTAE as the encoder underperformed, whereas those with SimSiam achieved more accurate groupings. This is likely because SimSiam is specifically trained to cluster similar objects closely in feature space, while ViTAE relies on a simple reconstruction loss, enhancing SimSiam's ability to densely cluster similar objects and distinctly separate different ones. Moreover, our analysis revealed that the most successful

TABLE I
COMPARISON OF PERFORMANCE AND TIME REQUIRED BETWEEN THE BASELINE AND THE PROPOSED METHOD

Method				Time [min]	AP-Worker			AP-handpallet			AP-Total		
Segmentation	Encoding	Clustering			AP	AP50	AP75	AP	AP50	AP75	AP	AP50	AP75
RAFT	SimSiam	UMAP	K-Means	8.97	**30.115**	**63.269**	22.073	**16.103**	**34.205**	**12.851**	**23.109**	**48.737**	**17.462**
			DBSCAN	23.8	27.341	56.780	22.120	13.951	32.781	9.280	20.646	44.780	15.700
		PCA	K-Means	**7.81**	28.888	57.592	23.359	13.776	30.411	9.257	21.332	44.001	16.308
			DBSCAN	-	-	-		-	-		-	-	
	ViTAE	UMAP	K-Means	20.5	29.498	60.357	23.081	11.325	23.202	9.033	20.412	41.779	16.057
			DBSCAN	-	-	-		-	-		-	-	
		PCA	K-Means	17.1	29.917	59.321	**25.096**	12.169	31.289	8.856	21.043	45.305	16.976
			DBSCAN	-	-	-		-	-		-	-	
Manual				626	37.741	67.497	36.478	24.917	51.558	21.316	31.329	59.528	28.897

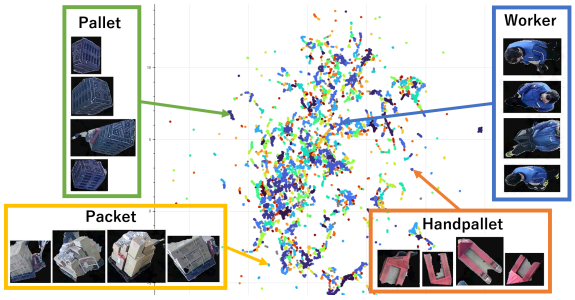

Fig. 10. 2D Mapping of the embedded objects

Fig. 11. Example of the segmentation results containing multiple semantics

Fig. 12. Example of the segmentation in the warehouse using SAM

accuracy outcomes were obtained with configurations incorporating SimSiam, underscoring its pivotal role in our proposed method. Additionally, configurations requiring less than 10 minutes predominantly included SimSiam as the encoder. This efficiency is attributed to SimSiam's high mapping accuracy in feature space, which reduced the time needed for noise data removal in our method's labeling process. Lastly, we observed that configurations using DBSCAN for clustering, even with SimSiam as the encoder, necessitated longer processing times. This is due to DBSCAN's characteristic of forming clusters without predefined limits, leading to a greater number of clusters and consequently, an increased workload of annotation task.

C. Analysis of Clustering Results

To check the distribution of clustered objects, we compressed the embedding representation using UMAP with an output dimension of 2 and performed 2D mapping and clustering using DBSCAN as Fig. 10 shows. We also selected four classes from it and displayed representative objects belonging to those classes and the label that should be assigned to the class. The proposed method maps similar objects to similar embedding representations and successfully creates clusters of objects that are believed to belong to the same class. We can complete labeling of all objects belonging to each cluster by labeling the cluster just once. We believe this feature of the proposed method significantly contributes to reducing the workload for annotation.

D. Issues with Segmentation

Our method struggles with multi-semantic segmentation, causing unclear labels as in Fig. 11, like "worker+packet" or "worker+handpallet" scenarios. In optical flow based object segmentation, challenges arise when multiple objects closely move together, such as in scenarios like "a worker in a warehouse with a handpallet". A straightforward solution involves discarding segmentation results with mixed semantics and generating new segmentations until required annotations are obtained, leveraging the fully automatic execution of our proposed method to avoid additional costs. Another limitation of our method is segmenting only moving objects. We propose integrating with models like Segment Anything Model (SAM) [30], illustrated in Fig. 12, which segments both moving and stationary objects and can be improved by parameter tuning or method combinations.

X. CONCLUSION

In this paper, we addressed digitalization challenges in logistics warehouses by developing a semi-automated framework. Our approach involved setting up an extensive camera network to enhance data collection, focusing on capturing object movements, analyzing truck berth operations, and utilizing instance segmentation for berth utilization assessment. The proposed semi-automated digitalization framework utilizes optical flow for object segmentation and representation learning to significantly reduce the annotation workload. The semi-automatically trained worker and handpallet recognition model showed lower accuracy compared to manually annotated datasets, but it expedited model creation by 98.6%.

ACKNOWLEDGMENT

This paper is based on results obtained from projects JPNP23003 and JPNP23025 commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This paper is partially supported by TR-USCO Nakayama Corp, JST KAKENHI(22K18422) and JST CREST(JPMJCR22M4).

REFERENCES

[1] Hui Sun and Xue Hao Gao. Research on the Optimization of Warehouse Logistics Efficiency Based on Order Sequencing. In *Proceedings of the 2022 10th International Conference on Information Technology: IoT and Smart City*, ICIT '22, pp. 274–279, New York, NY, USA, 2023. Association for Computing Machinery.

[2] Xiangwei Gong. Optimization Algorithm of Logistics Warehousing and Distribution Path based on Artificial Intelligence Technology. In *2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE)*, pp. 371–375, 2022.

[3] Ruozhen Qiu, Yue Sun, and Minghe Sun. A Robust Optimization Approach for Multi-Product Inventory Management in a Dual-Channel Warehouse under Demand Uncertainties. *Omega*, Vol. 109, p. 102591, 2022.

[4] Barbara Rita Barricelli, Elena Casiraghi, and Daniela Fogli. A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications. *IEEE Access*, Vol. 7, pp. 167653–167671, 2019.

[5] Yuto Fukushima, Yusuke Asai, Shunsuke Aoki, Takuro Yonezawa, and Nobuo Kawaguchi. DigiMobot: Digital Twin for Human-Robot Collaboration in Indoor Environments. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 55–62, 2021.

[6] Lixiang Zhang, Yan Yan, Yaoguang Hu, and Weibo Ren. Reinforcement Learning and Digital Twin-based Real-Time Scheduling Method in Intelligent Manufacturing Systems. *IFAC-PapersOnLine*, Vol. 55, No. 10, pp. 359–364, 2022. 10th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2022.

[7] Jiewu Leng, Douxi Yan, Qiang Liu, Hao Zhang, Gege Zhao, Wei Lijun, Ding Zhang, Ailin Yu, and Xin Chen. Digital Twin-Driven Joint Optimisation of Packing and Storage Assignment in Large-Scale Automated High-Rise Warehouse Product-Service System. *International Journal of Computer Integrated Manufacturing*, 2019.

[8] Xiaokang Zhou, Xuesong Xu, Wei Liang, Zhi Zeng, Shohei Shimizu, Laurence T. Yang, and Qun Jin. Intelligent Small Object Detection for Digital Twin in Smart Manufacturing With Industrial Cyber-Physical Systems. *IEEE Transactions on Industrial Informatics*, Vol. 18, No. 2, pp. 1377–1386, 2022.

[9] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment Everything Everywhere All at Once, 2023.

[10] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2955–2966, Los Alamitos, CA, USA, 2023. IEEE Computer Society.

[11] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7010–7021, 2022.

[12] Jiahao Lu, Chong Yin, Oswin Krause, Kenny Erleben, Michael Bachmann Nielsen, and Sune Darkner. Reducing Annotation Need in Self-explanatory Models for Lung Nodule Diagnosis. In Mauricio Reyes, Pedro Henriques Abreu, and Jaime Cardoso, editors, *Interpretability of Machine Intelligence in Medical Image Computing*, pp. 33–43, Cham, 2022. Springer Nature Switzerland.

[13] Nathan Elangovan, Ricardo V. Godoy, Felipe Sanches, Ke Wang, Tom White, Patrick Jarvis, and Minas Liarokapis. 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)*, pp. 11329–11335, 2023.

[14] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, Medical Image Computing and Computer Assisted Intervention − MICCAI 2017, pp. 399–407, Cham, 2017. Springer International Publishing.

[15] Donggeun Yoo and In So Kweon. Learning Loss for Active Learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 93–102, 2019.

[16] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M. Alvarez. Active Learning for Deep Object Detection via Probabilistic Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10264–10273, 2021.

[17] Guodong Shao and Moneer Helu. Framework for a Digital Twin in Manufacturing: Scope and Requirements. *Manufacturing Letters*, Vol. 24, pp. 105–107, 2020.

[18] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, 2023.

[19] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-Centric Sort: Rethinking Sort for Robust Multi-Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9686–9696, 2023.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[21] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software*, Vol. 3, No. 29, p. 861, 2018.

[22] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 226–231. AAAI Press, 1996.

[23] Gunnar Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis*, SCIA'03, pp. 363–370, Berlin, Heidelberg, 2003. Springer-Verlag.

[24] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pp. 402–419, Cham, 2020. Springer International Publishing.

[25] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[26] Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 91–99, Cambridge, MA, USA, 2015. MIT Press.

[29] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. (Accessed : 2024-02-01).

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. *arXiv:2304.02643*, 2023.