The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop “Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine”
PSBB25 , 9–11 June 2025, Moscow, Russia

Using the Segment Anything Model to Develop Control Pallet Loading System

Nikita Andriyanov[1], Svetlana Mikhailova[1], Xenin Fao[2]
[1]Financial University under the Government of the Russian Federation
[2]Huazhong University of Science and Technology, China

Keywords: Segmentation, Automatic Machine Learning, Computer Vision, Pallet Loading, Segment Anything Model.

Abstract

Modern warehouse complexes face the need for efficient and accurate pallet loading control in conditions of high dynamics and variety of objects. This paper proposes an approach to solving this problem based on the Segment Anything model (SAM) for automatic image tagging and the YOLOv8 model for subsequent accurate segmentation. This combination provides both high processing speed and adaptability to changing lighting conditions, partial overlaps, and complex object geometry. The proposed algorithm tracks changes in the area of segmented zones in order to estimate the addition of new cargo. The experiments show that YOLOv8 provides the best balance between accuracy and performance (Dice = 0.88), outperforming Mask R-CNN and the newer version YOLOv12. Additionally, the paper contains an analysis of the models' resistance to noise and visual distortions. The presented solution has the potential for integration into next-generation industrial logistics systems, reducing the need for manual annotation and increasing the autonomy of loading control.

1. Introduction

Modern logistics facilities today face escalating demands for speed, precision, and adaptability in the handling, storage, and dispatch of goods. Contemporary automation challenges in warehouse operations have risen to the point where traditional manual oversight can no longer keep pace with the volume and variability of incoming shipments. Among these critical tasks, the accurate monitoring and control of pallet loading stands out as one of the most labor-intensive and time-consuming processes. In high-throughput environments – in which hundreds or even thousands of cartons, boxes, and packages traverse conveyor belts and storage racks every hour—any delay or error in properly arranging and accounting for palletized freight can propagate downstream, causing costly misplacements, shipment delays, and logistical bottlenecks.

In this context, computer vision has emerged as a transformative enabler of next-generation warehouse intelligence. By leveraging advanced cameras, edge-computing hardware, and deep-learning algorithms, modern vision systems can automatically recognize, classify, and spatially localize individual items in real time. Such systems are capable of discerning an object’s shape, dimensions, orientation, and even surface characteristics–enabling not only binary “present/absent” checks but also volumetric assessments and overlap detection. The ability to process high-resolution video streams at frame rates exceeding 30 frames per second means that a vision-based controller can keep continual tabs on fast-moving forklifts, robotic palletizers, and human operators alike. This level of automation drastically reduces reliance on manual inspection, cuts down human error, and frees up personnel to focus on exception handling rather than routine snapshot verifications.

Against this backdrop of surging throughput requirements and zero-tolerance for errors, the need for sophisticated, scalable computer-vision and machine-learning solutions is more urgent than ever. In our work, we present an innovative, two-stage approach to precise pallet-loading control that brings together the state-of-the-art Segment Anything Model (SAM) for rapid, high-quality data annotation and the YOLO (You Only Look Once) framework for swift, accurate object segmentation. SAM’s breakthrough capability–powered by a massive Transformer backbone trained on over a billion masks allows for zero-shot extraction of object outlines from various visual prompts with minimal human intervention. These rich, automatically generated annotations dramatically reduce the dataset preparation overhead typically required to train supervised models.

Once an initial annotation scaffold is in place, the YOLOv8 architecture takes the baton, delivering fast, frame-by-frame object detection and mask generation optimized for industrial environments. Its single-stage pipeline, enhanced by Cross-Stage Partial (CSP) modules and anchor-free predictions, processes each frame in milliseconds, making it ideally suited for real-time deployment on edge devices. By integrating SAM’s annotation throughput with YOLO’s inference speed, our system forms a robust, end-to-end vision stack that can reliably track incremental additions to a pallet, even under shifting lighting conditions, partial occlusions, and varying camera angles. This combined SAM–YOLO paradigm thus establishes a new benchmark for autonomous pallet-loading control in dynamic logistics settings, delivering both the flexibility required for rapid model adaptation and the performance necessary for continuous, unattended operation.

A distinctive feature of the proposed method is a unique algorithm for analyzing the dynamics of changes in the segmented area using a top-mounted camera. This algorithm incorporates a specialized mathematical framework that accounts for the addition of new objects onto the pallet. The solution demonstrates high practical potential, delivering not only precise segmentation but also adaptability to varying conditions such as lighting changes or partial object occlusions.

Thanks to the universal architectures of SAM and YOLO, the developed system can scale and adapt to different logistics configurations, including autonomous warehouse robots, sorting

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-19-2025 | © Author(s) 2025. CC BY 4.0 License.

lines, and automated platforms. The method’s effectiveness has been empirically validated, and its stability and scalability pave the way for integration into next-generation logistics systems.

2. Related works

Recent advances in computer vision have had a profound impact on the automation of logistics processes, particularly in tasks such as object segmentation and pallet-loading state monitoring. Classical segmentation architectures—such as Mask R-CNN (He, 2017) and U-Net (Ronneberger, 2015)—laid the groundwork for precise object delineation, but their reliance on extensive manual annotation remains a significant bottleneck, especially in fast-moving, high-variability industrial settings. Standardized datasets like COCO (Lin, 2014) and LVIS (Gupta, 2019) have partially alleviated the annotation burden by providing large-scale, richly labeled images; however, highly specialized applications—such as real-time pallet-loading control—demand tailored solutions that account for the irregular shapes of cargo, complex overlap patterns, and rapid lighting fluctuations.

Several studies have explored the direct application of segmentation techniques in logistics. For instance, article (Zhang, 2020) employed YOLOv4 for pallet detection, while article (Wang, 2022) introduced a multi-view 3D reconstruction pipeline to estimate cargo volume. Although these methods achieve promising accuracy, they often struggle with real-time adaptability and fail to generalize across diverse warehouse layouts.

The advent of foundation models like SAM (Segment Anything Model) (Kirillov, 2023) marked a turning point by enabling zero-shot segmentation with minimal human input. Recent work by Chen (2023) confirmed SAM’s potential in industrial object detection, yet its integration into dedicated pallet-monitoring systems remains underexplored.

A particularly important research direction focuses on approaches that combine automated annotation, robust segmentation, and adaptive algorithms to handle geometric heterogeneity of palletized loads. For example, authors (Li, 2021) demonstrated a depth-sensor–based method, but it exhibited limited scalability under variable illumination and overlapping scenarios. Continuous innovation is needed to bridge these gaps and deliver truly resilient, end-to-end segmentation pipelines for modern logistics environments.

Specialized tools of using generative models to improve detectors and segmentators quality are provided in work (Andriyanov, 2024).

Methods for 3D-structure analysis (Andriyanov, 2021; Sun, 2024) offer valuable insights into the geometric properties of palletized loads but are not always suitable for real-time applications due to their computational complexity and reliance on specialized depth sensors. Consequently, there remains a substantial gap in existing solutions that can simultaneously deliver high processing speed, segmentation accuracy, and adaptability to changing warehouse conditions. The present work bridges this gap by leveraging the rapid, high-quality annotation capabilities of SAM, the real-time inference speed and precision of YOLO, and a proprietary algorithm that dynamically tracks changes in object surface area on the pallet.

Moreover, there is growing interest in multimodal systems that combine visual input with textual or semantic prompts, enabling the segmentation model to interpret not only the shape and position of items but also their category and handling requirements. Recent advances in zero-shot segmentation frameworks—such as SAM and SEEM—demonstrate the feasibility of fully automated model adaptation to the unique appearance and arrangement of goods in a given warehouse, without costly manual label creation or extensive retraining. This combined visual-semantic prompting approach holds promise for creating next-generation logistics platforms that can self-optimize based on real-world operating conditions.

In summary, the evolution of computer vision techniques has dramatically advanced the state of the art in logistics automation, yet critical challenges remain in unifying high throughput, segmentation fidelity, and operational robustness. Classical two-stage networks such as Mask R-CNN and encoder–decoder models like U-Net paved the way for precise object delineation but continue to impose prohibitive annotation overheads in dynamic warehouse environments. While large-scale benchmarks (COCO, LVIS) have mitigated this burden, they fall short of capturing the idiosyncratic geometries, occlusions, and lighting variations characteristic of palletized cargo. Early industrial adaptations – ranging from YOLOv4-based pallet detectors to multi-view 3D reconstructions – have demonstrated feasibility yet often lack real-time adaptability and cross-site generalizability.

The introduction of foundation models such as SAM, capable of zero-shot segmentation, represents a pivotal paradigm shift by drastically reducing manual labeling requirements without sacrificing mask quality. Nevertheless, the seamless integration of such models into end-to-end pallet-monitoring pipelines remains nascent. Similarly, depth-sensor-driven and 3D-analytic approaches offer valuable geometric insight but struggle to meet the stringent demands of real-time throughput.

Our synthesis emphasizes the urgent need for hybrid frameworks that leverage automated, high-quality annotation (SAM), ultra-fast inference (YOLO), and adaptive algorithms to dynamically track load composition. Moreover, the emerging paradigm of multimodal prompting – combining visual and semantic cues – holds promise for contextualizing segmentation outputs with cargo metadata, further reducing human intervention. Moving forward, research should prioritize scalable architectures that natively support zero-shot adaptation, real-time performance, and multimodal integration, thereby enabling truly autonomous, resilient pallet-loading control systems for tomorrow’s smart warehouses.

3. Materials and methods

In this study, we employed authentic video recordings of forklift operations under real warehouse conditions. The total amount of raw footage amounted to approximately 20 hours; from this, we selected a representative subset of 20 minutes that covered the most critical pallet-loading scenarios, including peak throughput periods and challenging loading angles. To ensure a consistent temporal annotation framework, the continuous video stream was converted into individual image frames at a rate of one frame per second, yielding a dataset of 1,200 frames—providing extra overlap for late-stage validation.

These frames encompass a wide variety of operational circumstances: fluctuating illumination levels caused by overhead lighting changes, partial and full object occlusions when pallets overlap or are stacked, and variations in camera viewing angles as forklifts approach from different directions.

For automatic object segmentation within these frames, we utilized the FastSAM (Fast Segment Anything Model) tool from the Ultralytics library. FastSAM is a pre-trained model based on the YOLOv8 architecture that generates precise object masks in response to approximate rectangular prompts (bounding boxes). This approach dramatically reduces manual annotation effort compared to fully hand-labeled datasets, while still delivering high-quality segmentation.

A key advantage of FastSAM lies in its robustness to imprecise inputs: even when bounding-box prompts are only roughly drawn, the model accurately infers object contours. This capability is especially important for dynamic scenes featuring partial overlaps, complex pallet geometries, and rapid changes in object position, enabling reliable mask extraction with minimal user intervention.

Figure 1 shows the original frame on the left and the frame marked up with FastSAM on the right.

Figure 1. Examples of source and labeled images

Figure 2 demonstrates segmented data.

Figure 2. Segmented objects

The core SAM model (Segment Anything Model) is a universal image segmentation algorithm designed to extract object boundaries using various forms of input prompts – including points, bounding boxes, and even textual descriptions – without requiring additional training or fine-tuning for new tasks. This flexibility makes SAM particularly powerful in environments where data diversity is high and manual annotation is costly or impractical.

Built upon a scalable and modular Transformer-based architecture, SAM was trained on the massive SA-1B dataset, which contains over one billion high-quality segmentation masks spanning a wide range of object types, shapes, and contexts. This extensive training corpus enables the model to generalize effectively across unseen scenarios, making it suitable not only for academic benchmarks but also for deployment in complex, real-world environments.

The model supports both automatic segmentation – where masks are generated without any human interaction – and interactive segmentation, which allows users to refine results by providing feedback or targeted input cues. This dual-mode functionality provides flexibility depending on the precision and speed requirements of the use case. SAM demonstrates impressive robustness to environmental noise, visual clutter, object occlusions, and complex backgrounds, making it well-suited for industrial applications.

A particularly critical advantage of SAM lies in its ability to function effectively in low-data or zero-shot settings, where traditional models would typically require extensive retraining or large labeled datasets. This makes SAM a strong candidate for industrial monitoring tasks such as pallet-loading control, where scenes are dynamic, lighting conditions vary, and the need for real-time, high-accuracy segmentation is paramount. Its ability to deliver precise object boundaries without reliance on task-specific training pipelines represents a major leap forward in automating visual intelligence across logistics and manufacturing workflows.

4. Results and Discussion

The primary evaluation metric used in this study was the Dice coefficient, also known as the Dice Similarity Coefficient (DSC). This metric is widely adopted in segmentation tasks, particularly in scenarios where the overlap between predicted and ground truth regions must be accurately quantified.

In the context of pallet-loading control, precise object boundary detection is essential, as even small segmentation errors can lead to incorrect object counts, misinterpretation of loading completeness, or failure to detect overlapping packages. The Dice metric provides a direct measure of similarity between the predicted segmentation mask and the actual (reference) mask, making it highly suitable for evaluating performance in such spatially sensitive tasks.

Formally, the Dice coefficient is defined as:

$$Dice = \frac{2\,|\,A \cap B\,|}{|\,A\,| + |\,B\,|}, \tag{1}$$

where $\boldsymbol{A}$ is the set of predicted mask pixels, and $\boldsymbol{B}$ is the set of ground truth mask pixels. The metric ranges from 0 (no overlap) to 1 (perfect match), and emphasizes both precision and recall simultaneously.

Given the operational need for high segmentation accuracy under varying lighting, angles, and object occlusions in warehouse conditions, the Dice score is an appropriate and reliable metric for assessing the real-world applicability of our segmentation approach. Its sensitivity to both over-

segmentation and under-segmentation allows us to ensure that the proposed system performs robustly in dynamic industrial environments.

Table 1 shows the comparison results of different approaches.

Model	Dice
YOLOv12	0.86
YOLOv8	0.88
Mask R-CNN	0.84

Table 1. Comparison of Image Segmentation.

The YOLOv8 model achieved the highest segmentation accuracy in our evaluation, with a Dice score of 0.88, outperforming both YOLOv12 (0.86) and the more traditional two-stage Mask R-CNN (0.84). This superior performance of YOLOv8 can be attributed to several architectural enhancements introduced in this version, such as the integration of Cross Stage Partial (CSP) connections, decoupled head structures, and optimized anchor scaling mechanisms. These improvements not only enhance feature reuse and gradient flow but also lead to more stable and precise object boundary predictions, particularly in crowded and variable industrial scenes.

Despite being a newer iteration, YOLOv12 underperforms slightly compared to YOLOv8 in terms of segmentation accuracy. This may indicate that YOLOv12 has been optimized primarily for speed or model size – potentially through lightweight backbones or transformer-like modules – possibly at the cost of fine-grained mask quality. Further investigation is required to determine the extent to which these architectural trade-offs affect detection robustness, especially in cluttered or occluded environments like pallet stacks.

Mask R-CNN, while once a benchmark in object segmentation, shows the lowest performance among the compared models in our setup. This is likely due to its inherently slower, two-stage architecture, which separates region proposal from mask prediction. In dense logistic scenarios, such as those involving overlapping boxes or varying lighting conditions, Mask R-CNN may accumulate error across stages, leading to reduced segmentation fidelity and degraded spatial accuracy.

From an industrial deployment perspective, YOLOv8 offers the most practical balance between precision, speed, and implementation complexity. Its high segmentation quality combined with real-time inference capability makes it ideal for on-device execution in warehouse automation systems. However, the potential inference speed advantages of YOLOv12 (e.g., frames per second, memory efficiency) warrant further benchmarking under production conditions.

To further enhance system performance and robustness, we recommend integrating additional evaluation metrics such as Intersection over Union (IoU), pixel-wise accuracy, and false positive/negative rates. Moreover, qualitative tools such as error heatmaps, boundary analysis, and visualization of common failure cases (e.g., object merging or incomplete segmentation) will be crucial for refining the segmentation pipeline and ensuring reliable operation in diverse logistics environments.

To assess the robustness of the proposed algorithm under realistic and potentially challenging operating conditions, an additional experiment was conducted involving the intentional introduction of noise distortions. The objective was to simulate environmental variations that commonly occur in warehouse settings, such as sensor interference, inconsistent lighting, and partial occlusion from nearby objects.

As part of this experiment, modified versions of the test dataset were generated by applying three types of controlled visual perturbations. First, Gaussian noise with varying standard deviations was added to simulate sensor-level interference and low-quality video feeds. This type of noise helps evaluate the model’s ability to maintain segmentation accuracy when the input data becomes visually degraded. Second, brightness adjustments—both increases and decreases—were applied to individual frames to replicate real-world fluctuations in illumination, such as those caused by moving forklifts casting shadows or temporary light obstructions. Third, artificial shadow overlays and occlusion patterns were introduced to simulate the presence of overlapping objects or the inconsistent visibility of palletized items under dynamic conditions.

By subjecting the model to these distortions, we aimed to evaluate not only its base-level segmentation performance but also its resilience to common operational noise, which is critical for reliable deployment in industrial environments where visual clarity cannot always be guaranteed.

Table 2 presents the results of experiment.

Condition	Dice YOLOv8	Dice FastSAM
No distortions	0.88	0.85
Gaussian noise ($\sigma = 20$)	0.87	0.83
Brightness variation (±25%)	0.86	0.82
Shadows and partial occlusion	0.84	0.78

Table 2. Segmentation Accuracy under Visual Distortions.

The experiment demonstrates that while FastSAM maintains relatively stable segmentation performance under moderate noise and lighting variations, it exhibits noticeable sensitivity when subjected to strong occlusions and overlapping objects. This indicates that although FastSAM performs well in clean or lightly distorted scenes, its segmentation quality tends to degrade in more complex visual scenarios typical of dynamic warehouse environments—particularly when multiple packages overlap or when shadows obscure clear object boundaries. The decline in performance under partial occlusion suggests a limitation in the model’s ability to infer complete object masks when only fragmented visual cues are available.

In contrast, YOLOv8 consistently preserves high segmentation accuracy across all tested distortion scenarios, including under the influence of Gaussian noise, changes in brightness, and artificially introduced shadows. This robustness highlights the strength of YOLOv8’s feature extraction capabilities, likely due to its optimized architecture and training strategies that enhance generalization. Notably, even in scenes with reduced visual clarity, YOLOv8 manages to retain structural integrity in its predicted masks and continues to accurately delineate individual object boundaries.

Such performance under stress conditions is especially valuable in industrial and warehouse settings, where visual inconsistencies—caused by camera motion, variable ambient lighting, or occlusions from equipment and workers—are the norm rather than the exception. The results affirm that YOLOv8

is well-suited for real-world deployment in logistics operations, offering a reliable foundation for automated pallet monitoring and item tracking, even when ideal imaging conditions cannot be guaranteed. This resilience makes it a preferable choice for systems that must operate continuously and with minimal human supervision.

Due to the fact that we choose the area of interest with segmented area counts larger than the pallet, we observe the area drop when the loader takes a new object, as shown in Figure 3.

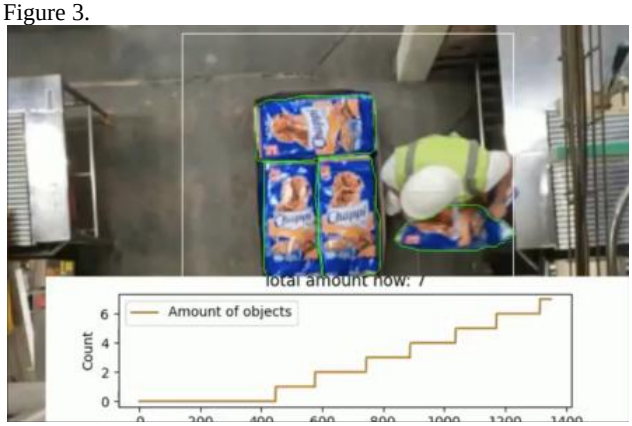

Figure 3. Object number estimation

Building on the observation provided in Figure 3, the segmentation-based area tracking produces a distinctive stepwise pattern over time, which serves as a powerful indicator of discrete loading events. Because the region of interest (ROI) in the segmentation process is deliberately defined to include all objects placed on top of the pallet—regardless of shape or position – each new addition by the loader results in a measurable increase in the total segmented area. Conversely, when a loader temporarily obstructs or adjusts an object, or when repositioning occurs, we may see small fluctuations. However, the most significant and consistent trend is the sharp increase in area each time a new object is added.

This behavior forms a staircase-like curve, where each horizontal plateau corresponds to a static state – i.e., a moment when no new items are added – while each upward step signals the placement of a new object. This temporal profile allows for an implicit count of objects added, without relying on direct object recognition or classification. The method thus enables object-level event detection purely through geometric changes in the segmented mask area, which proves especially useful in real-time monitoring contexts.

Importantly, a sharp drop in the segmented area to near-zero levels acts as a reliable signal that the pallet has been removed or cleared from the field of view, typically due to its transport out of the camera’s coverage zone. This behavior is crucial for resetting the counting process: rather than relying on external triggers or human input, the system autonomously determines when a new loading cycle should begin. This automated reset mechanism minimizes error accumulation between cycles and enhances the system’s autonomy in multi-cycle operations.

Overall, Figure 3 not only confirms the system’s sensitivity to discrete loading actions but also demonstrates its ability to segment operational sequences into logical units: loading events, static holds, and pallet departure. This segmentation timeline can be further integrated into downstream analytics – such as efficiency tracking, anomaly detection, or worker performance evaluation – making it a foundational element of intelligent warehouse automation.

5. Conclusions

The proposed pallet loading segmentation and control system, based on a combination of SAM and YOLOv8 models, demonstrates high accuracy, adaptability and practical applicability in warehouse logistics. The developed area dynamics analysis algorithm allows tracking loading events in real time and opens up prospects for integration into autonomous logistics complexes.

The integrated use of zero-shot segmentation from SAM for automated annotation and high-speed, accurate inference from YOLOv8 allows for the formation of a stable architecture suitable for implementation in industrial processes without the need for lengthy manual adjustment. The experiments confirmed that the system maintains stable operation in the presence of visual distortions, such as noise, changing lighting and partial overlapping of objects, which is critical for real operating conditions. In addition, the ladder profile of the area dynamics of segmented objects provides a simple and reliable metric for counting loaded elements, allowing the system to autonomously identify the completion of a loading cycle and begin a new one.

The paper also highlights the potential of multimodal and adaptive approaches that can extend the capabilities of the current architecture by integrating semantic hints and self-correction mechanisms. All this opens the way to creating more versatile, self-adjusting solutions in the field of computer vision for logistics, capable of functioning in conditions of high variability and minimal human intervention.

Thus, the proposed system can become an important step towards fully automated warehouses of the new generation, providing not only loading control, but also intelligent analysis of logistics processes in real time.

References

Andriyanov, N., Dementiev, V., Kondratiev, D., 2021: Tracking of Objects in Video Sequences. Intelligent Decision Technologies: Proceedings of the 13th KES-IDT 2021 Conference, 253-262.

Andriyanov, N., Kim, A., Fao X., 2024 Using Generative Models to Improve Fire Detection Efficiency. 2024 X International Conference on Information Technology and Nanotechnology (ITNT), Samara, Russian Federation, 2024, pp. 1–4, doi: 10.1109/ITNT60778.2024.10582386.

Chen, J., Li, Y., Zhang, Q., Wang, L., 2023: SAM for Industrial Object Detection: A Case Study in Warehouse Automation. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

Gupta, A., Dollar, P., Girshick, R., 2019: LVIS: A Dataset for Large Vocabulary Instance Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017: Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980-2988.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., 2023: Segment Anything. arXiv preprint arXiv:2304.02643.

Li, H., Zhang, Z., Liu, Y., Wang, J., 2021: Depth-Based Pallet Loading Analysis in Dynamic Environments. IEEE Sensors Journal, 21(15), 17281-17290.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017: Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980-2988. doi.org/10.1109/ICCV.2017.324.

Ronneberger, O., Fischer, P., Brox, T., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234-241.

Sun, X., He, L., Jiang, H., Li, R., Mao, W., Zhang, D., Majeed, Y., Andriyanov, N., Soloviev, V., Fu, L., 2024: Morphological Estimation of Primary Branch Length of Individual Apple Trees During the Deciduous Period in Modern Orchard Based on PointNet++. Computers and Electronics in Agriculture, 220, 108873. doi.org/10.1016/j.compag.2024.108873.

Wang, L., Zhang, Y., Chen, Z., Liu, J., 2022: 3D Volume Estimation for Logistics Cargo via Multi-View Fusion. IEEE Transactions on Instrumentation and Measurement (TIM), 71, 1-12.

Zhang, Y., Wang, C., Li, X., 2020: Real-Time Pallet Detection Using YOLOv4 in Automated Warehouses. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 10157-10163.