Safety-Centric Monitoring of Structural Configurations in Outdoor Warehouse Using an UAV

Assia Belbachir[1][a], Antonio M. Ortiz[1][b], Ahmed Nabil Belbachir[1][c] and Emanuele Ciccia[2]
[1]*NORCE Research AS, Grimstad, Norway*
[2]*ABS - Acciaierie Bertoli Safau S.p.A., Udine, Italy*

Keywords: Industrial Safety, Computer Vision, Warehouse Management, Geometric Reasoning, Steel Bar Manufacturing, Segment Anything Model, UAV.

Abstract: In industrial warehouse environments, particularly in steel bar manufacturing scenarios, ensuring the structural stability of stacked bars is essential for both worker safety and operational efficiency. This paper presents a novel vision-based framework for automatic safety validation of outdoor storage bays using a dual-resolution implementation of the Segment Anything Model (SAM). The system processes video streams coming from drone (AUV) by combining zero-shot segmentation with geometric reasoning to assess lateral and frontal support conditions in real time. At each frame, SAM is applied at two scales to extract both fine-grained support components and large bulk regions. A morphological proximity rule reclassifies unsupported regions based on contact with multiple smaller support masks. Additionally, a frontal-view analysis computes bar-end centroids and applies a triangle-based inclusion test to determine correct placement. Experimental results on real warehouse videos demonstrate robust safety classification under occlusion and clutter, with interactive frame rates and no need for manual annotation. The proposed framework offers a lightweight, interpretable solution for automated safety monitoring in complex industrial environments.

1 INTRODUCTION

The rise of Industry 4.0 has led to the widespread adoption of computer vision systems in manufacturing and logistic workflows, allowing automation in areas such as defect inspection, dimensional metrology, and human–machine interaction monitoring for improved throughput and safety (Smith and Lee, 2019). In parallel, logistics and warehousing operations increasingly rely on vision systems for inventory tracking, object localization, and robot guidance (Patel and Gupta, 2020). In industrial environments, such as steel bar manufacturing facilities, improper stacking or insufficient bracing of materials poses serious risks, including potential collapses, equipment damage, and workplace injuries.

Despite the severity of hazards, structural stability assessments remain predominantly manual, making them prone to human error, subjective interpretation, delayed response, and inconsistent execution. This highlights a critical need for automated, vision-based solutions that can ensure reliable and timely safety validation in such high-risk environments.

Existing computer vision mechanisms for industrial safety monitoring often focus on detecting personnel, identifying personal protective equipment, or spotting unsafe behaviors. Meanwhile, segmentation-based solutions can localize and label individual objects with high accuracy, but often require task-specific training data and struggle with generalization in cluttered or outdoor scenes. Moreover, traditional reasoning techniques, while interpretable, lack robustness to occlusion and visual variability, making them insufficient when deployed in complex storage environments. Instance-level models such as Mask R-CNN achieve high accuracy in part segmentation, but demand extensive annotated datasets and exhibit brittleness under domain shifts (He et al., 2017). The recent Segment Anything Model (SAM) overcomes annotation bottlenecks by providing zero-shot, promptable mask proposals across domains without retraining (Kirillov et al., 2023), yet its single-scale outputs may under-segment small bracing elements or over-segment large bulk regions when deployed in

[a] https://orcid.org/0000-0002-1294-8478
[b] https://orcid.org/0000-0002-7145-8241
[c] https://orcid.org/0000-0001-9233-3723

Belbachir, A., Ortiz, A. M., Belbachir, A. N. and Ciccia, E.
Safety-Centric Monitoring of Structural Configurations in Outdoor Warehouse Using an UAV.
DOI: 10.5220/0013675300003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 2*, pages 471-478
ISBN: 978-989-758-770-2; ISSN: 2184-2809

isolation.

Complementary to learning-based segmentation, classical geometric reasoning techniques (e.g., Hough and RANSAC) detect primitives such as lines, circles, and triangles for structural analysis in construction and logistics applications (Duda and Hart, 1972). Handcrafted pipelines combining thresholding and shape tests can identify support wedges or triangular braces (Eiffert et al., 2021), but they lack robustness to visual clutter, occlusion, and lighting variability common in outdoor warehouses. More recent volumetric extensions using 3D radiance fields for support estimation (Cen et al., 2023), but incur prohibitive computational cost for real-time monitoring.

Multi-scale segmentation and proximity analysis represent a promising middle ground. Deep networks with feature pyramids capture both fine and coarse structures (Wu and Zhang, 2019), while morphological dilation and contact-based heuristics have been applied to validate part assembly in robotics (Zhang et al., 2022). To our knowledge, no existing approach unifies *zero-shot mask generation* at multiple resolutions with simple, interpretable geometric tests and proximity reclassification to deliver real-time stability checks of stacked materials in image streams.

In this work, we address these challenges with a novel vision-based safety monitoring framework for *safety validation* of outdoor steel-bar storage bays using top- and front-view images. Our contributions are:

- **Dual-Scale Zero-Shot Segmentation.** Using the Segment Anything Model (SAM) with lightweight geometric reasoning to assess structural support both from top and front-view images (`points_per_side`=32 and 64) to capture both fine support components and large bulk regions without any manual labeling (Kirillov et al., 2023).
- **Morphological Proximity Reclassification.** We introduce a lightweight dilation-based rule that reclassifies large, initially "at-risk" regions as *supported* only when contacted by at least three distinct fine-scale masks, ensuring interpretable, topology-aware decisions (Duda and Hart, 1972).
- **Triangle-Based Frontal Validation.** We extract bar-end centroids from front views and form a minimal support triangle to verify correct bar placement within safe boundaries, inspired by geometric support tests in logistics vision (Lee and Kim, 2021).
- **Real-World Warehouse Evaluation.** We demonstrate robustness and efficiency on outdoor manufacturing video streams—achieving a good frame rates and safety detection reliability compared to other approaches.

The proposed approach is efficient, generalizable across varying conditions, and suitable for real-time deployment in industrial settings.

The remainder of this paper is organized as follows. Section 2 reviews related work in industrial segmentation and safety monitoring. Section 3 formalizes our bay stability criteria. Section 4 details the proposed dual-SAM segmentation and geometric algorithms. Section 5 presents experimental results and performance analysis, and finally, Section 6 concludes with future directions.

2 RELATED WORK

Vision-based safety systems in manufacturing have primarily focused on human and equipment monitoring—detecting PPE compliance, unsafe actions, or machine faults (Smith and Lee, 2019) (Patel and Gupta, 2020). These approaches often neglect material stability issues, such as improperly braced stacked steel bars, which pose serious safety risks.

Semantic segmentation methods like Mask R-CNN (He et al., 2017) have shown high accuracy in part-level detection but require large annotated datasets and struggle with domain shifts. The Segment Anything Model (SAM) (Kirillov et al., 2023) enables zero-shot mask generation, greatly reducing annotation needs. However, its single-scale outputs can under-segment small supports or over-segment large objects in cluttered scenes.

Classical geometry-based techniques, including Hough and RANSAC (Duda and Hart, 1972) (Lee and Kim, 2021), have been used for detecting structural primitives. Hybrid pipelines combining segmentation and shape heuristics (Eiffert et al., 2021) or 3D volumetric reasoning (Cen et al., 2023) offer deeper structural insights, but often lack robustness or real-time efficiency.

Multi-scale segmentation (Wu and Zhang, 2019) and proximity analysis (Zhang et al., 2022) have been used in robotics to verify physical support, but existing work does not integrate zero-shot multi-scale segmentation with interpretable geometric reasoning for real-time stability validation.

Gap and Our Contribution. We address this gap by unifying dual-resolution SAM segmentation with morphological proximity rules and triangle-based geometric validation, enabling efficient and interpretable safety checks for stacked materials in industrial video streams.

Figure 1: Illustration of one Bay/box of outdoor steel bars.

3 PROBLEM DEFINITION

In steel bar manufacturing, storage areas (referred to as *bays* or *boxes*) are designated zones where steel bars are stacked and temporarily held before further processing or transportation. An example of a bay/box is shown in Figure 1. The structural stability of each bay is critical to ensure operational safety, as improperly supported stacks can lead to hazardous situations, including material collapse and injury. A bay is considered structurally **safe** when sufficient support is presented on both lateral sides and at the front, thus meeting the specific support criteria as follows:

- **Left Support:** At least two support structures are detected on the left side of the bay.
- **Right Support:** At least two support structures are detected on the right side of the bay.
- **Front Support:** The steel bars are positioned within a predefined virtual triangular region at the front of the bay. Bars located outside this region are considered improperly placed and can pose safety risks.

These support structures typically include physical components such as wedges or inclined bars that secure heavy loads. The virtual triangular region at the front serves as a spatial guide to define the correct placement of the steel bars, ensuring that they are adequately supported and do not extend beyond safe limits.

Formal To formalize this, we define the input as a video stream where each frame is represented as a color image $I \in \mathbb{R}^{H\times W\times 3}$, where H and W denote height and width, respectively. Within each frame, we selected a set of n predefined *bays* (or *boxes*), each denoted as $B_i \subset I$, where $i = 1,2,\ldots,n$. Each bay B_i must satisfy a set of specific structural safety conditions to be considered as *safe*. We have defined three support zones with respect to the spatial information of each bay:

- $\mathcal{L}(B_i)$: left support zone of bay B_i
- $\mathcal{R}(B_i)$: right support zone of bay B_i
- $\mathcal{F}(B_i)$: front support zone of bay B_i

Let $\mathcal{T}$ denote the set of all detected support structures in the image, i.e., $\mathcal{T} = T_j \subset I$. The number of support elements within each zone is then computed as follows:

$$N_L(B_i) = |\{T_j \in \mathcal{T} \mid T_j \subset \mathcal{L}(B_i)\}| \quad (1)$$

$$N_R(B_i) = |\{T_j \in \mathcal{T} \mid T_j \subset \mathcal{R}(B_i)\}| \quad (2)$$

$$N_F(B_i) = |\{T_j \in \mathcal{T} \mid T_j \subset \mathcal{F}(B_i)\}| \quad (3)$$

The binary *safety condition* for each bay B_i is defined as:

$$S(B_i) = \begin{cases} 1, & \text{if } N_L(B_i) \geq 2 \wedge N_R(B_i) \geq 2 \wedge N_F(B_i) \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

A value of $S(B_i) = 1$ indicates that the bay B_i meets all safety requirements, while $S(B_i) = 0$ flags it as potentially unsafe due to insufficient support structures.

4 FRAMEWORK OF THE PROPOSED APPROACH

This section provides an overview of our proposed vision-based safety validation framework, which is designed to assess the structural stability of material stacks in outdoor warehouse environments. The system leverages multi-resolution zero-shot segmentation and geometric reasoning to validate support conditions from both top and front camera views.

Figure 2 illustrates the overall architecture of our method, which consists of the following core components:

1. **Dual-View Video Acquisition:** The system captures synchronized video streams from two perspectives: a top-down view to assess lateral support conditions, and a frontal view to validate bar-end positioning.
2. **Multi-Resolution Segmentation with SAM:** Each frame is processed using the Segment Anything Model (SAM) at two different resolutions: a coarse scale (`points_per_side` = 32)

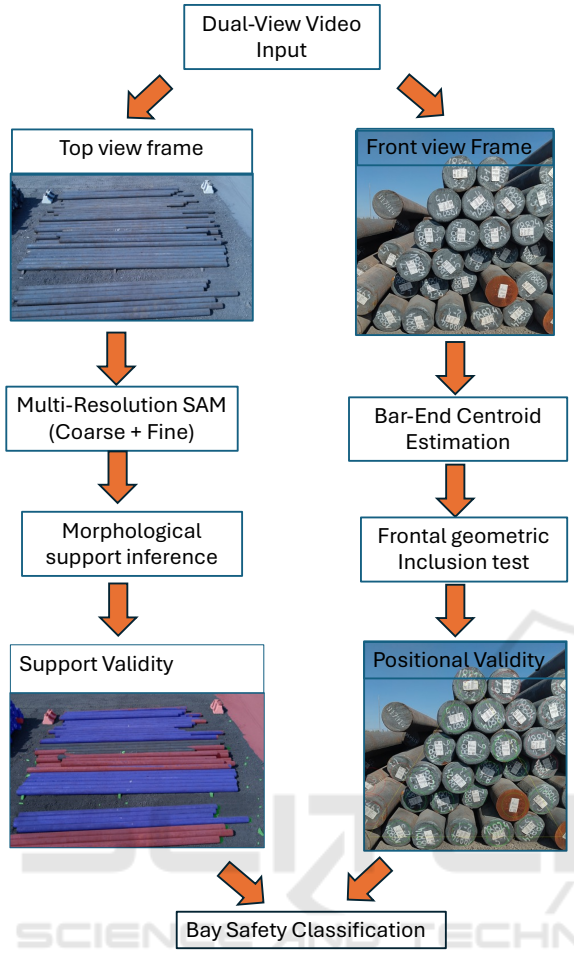


Figure 2: Illustration of the developed framework.

to segment large bulk materials, and a fine scale (`points_per_side` = 64) to detect smaller structural supports such as wooden braces, metallic beams, or narrow wedges.

3. **Morphological Support Inference:** In the top view, we apply a proximity-based rule: large masks are classified as *supported* if they are in direct contact with at least three smaller support masks. Contact is established using binary dilation and overlap checking, mimicking morphological reasoning rather than rigid geometry.
4. **Frontal Geometric Validation:** For frontal frames, we compute centroids of bar ends and apply a triangle-based inclusion test. The triangle is defined using warehouse-specific reference points, and each bar-end must fall within the triangle to be considered properly positioned and safe.
5. **Safety Classification and Visualization:** The system outputs a per-frame safety assessment, flagging any detected violations such as unsupported materials or improperly placed bars. Results are visualized in real time with overlaid masks and support indicators for operator feedback.

This modular pipeline ensures interpretability, scalability to new material types, and robust operation under cluttered or low-visibility conditions—all without the need for manual annotation or retraining.

4.1 Multi-Resolution SAM for Dual Views

To enable real-time safety validation of steel bar storage bays, we propose a vision-based algorithm that leverages Segment Anything Model (SAM) for automatic mask generation, combined with proximity-based geometric reasoning. The approach operates directly on individual video frames and is designed to identify structural support elements without requiring prior annotation or domain-specific retraining. The algorithm incorporates two key components (i) *top-view support detection*, which verifies lateral and rear support structures, and *front-view validation*, which assesses frontal bar placement using centroid-based geometric constraints. Each component uses SAM's zero-shot segmentation capability at multiple resolutions to extract both fine-grained and large-scale structural features, enabling robust performance under challenging visual conditions such as occlusion, clutter, and lighting variability.

4.2 Top-View Support Detection Algorithm

The top-view detection module processes each frame extracted from the input video and convert them from BGR to RGB format to meet the input requirements of the SAM framework. Two SAM-based automatic mask generators are used in parallel, each configured with a different resolution (specifically, *points_per_side* = *32* and *64*). This dual-resolution strategy enables the capture of multi-scale structural features within the frame.

The masks produced by both generators are aggregated and classified according to their pixel area. Masks falling within a predefined small-area range are interpreted as potential *support points*, while larger masks are considered *critical regions* that may require structural evaluation. Small-area masks are rendered in green, indicating supportive features, whereas large-area masks are initially colored red to denote potential risks. To determine the structural safety of the red regions, a proximity-based reclassi-

Algorithm 1: Dual SAM-Based Support Detection.

Require: Input video V
Ensure: Output video $\hat{V}$ with colored safety masks
Load SAM model with checkpoints
Initialize two SAM mask generators with $p = 32$ and $p = 64$
for each frame F in video V **do**
  Convert F from BGR to RGB
  Generate masks: $M_{32} \leftarrow \text{SAM}_{32}(F)$, $M_{64} \leftarrow \text{SAM}_{64}(F)$
  $M \leftarrow M_{32} \cup M_{64}$
  Separate masks into:
  - **Small masks** S: $a \in (0, 5000)$
  - **Large masks** L: $a \in [6000, 1.2 \times 10^6]$
  Label small masks as green, large as red
  for each red mask $r \in L$ **do**
    Dilate r to get dilated$_r$
    Count green masks $g \in S$ touching dilated$_r$
    if count ≥ 3 **then**
      Reclassify r as blue
    end if
  end for
  Overlay green, red, and blue masks onto F
  Blend mask overlay with original frame
  Write processed frame to $\hat{V}$
end for
Save $\hat{V}$ as output video

fication is performed. Each large red mask undergoes morphological dilation, and the algorithm checks for overlapping or nearby green regions. If a red region is in contact with at least three distinct green masks, it is reclassified as structurally supported and recolored blue. This proximity threshold ensures that only well-supported regions are marked safe. In the final step, mask overlays are combined with the original video frame using alpha blending to preserve visual context. Each reclassified (blue) region may also be annotated with the number of touching green masks for interpretability. The processed frames are then compiled into a new output video that visually communicates safety-related insights throughout the footage. This approach offers a semi-automated mechanism for identifying and verifying structural support in steel bar manufacturing environments, with potential applications in quality assurance, anomaly detection, and operator safety systems (see algorithm 1).

4.3 Front-View Safety Detection via SAM and Triangle Geometry

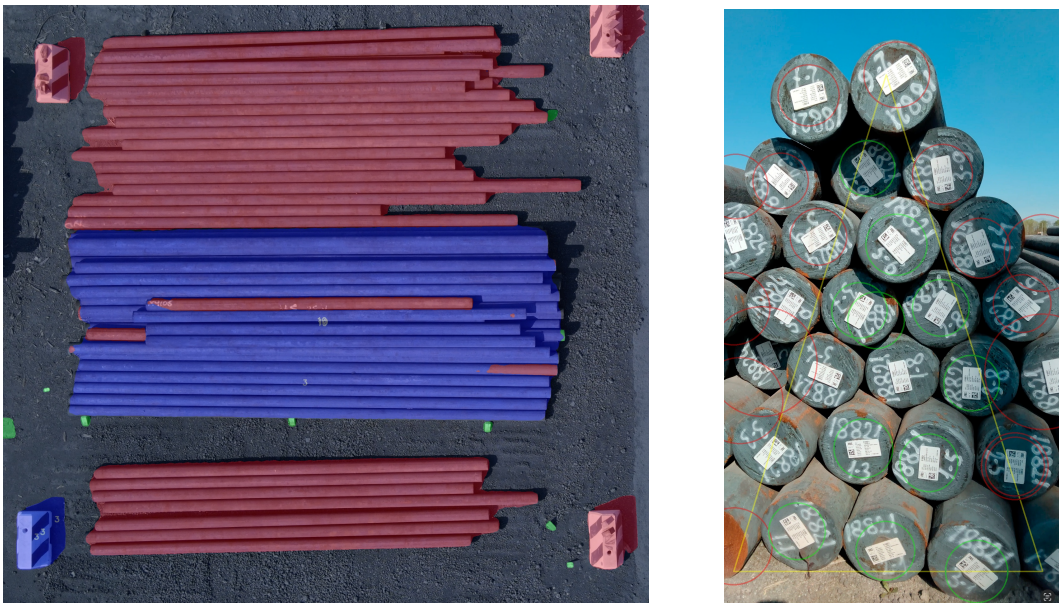
To evaluate frontal safety in steel bar configurations, we introduce a geometric reasoning algorithm that

Algorithm 2: Front safety detection via SAM and triangle geometry.

Input: Image I from front view, SAM model $\mathcal{M}$, thresholds $(A_{\min}, A_{\max}, \gamma)$
Output: Safety status (`SAFE` or `UNSAFE`) and annotated image
Convert I to RGB format
Generate mask set $\mathcal{S} \leftarrow \mathcal{M}(I)$
Initialize empty set of bar centers $\mathcal{C} \leftarrow \emptyset$
for each mask $s \in \mathcal{S}$ **do**
  Compute area a_s and contour c_s
  if $a_s \notin [A_{\min}, A_{\max}]$ **then**
    continue
  end if
  Compute circularity κ_s of c_s
  if $\kappa_s < \gamma$ **then**
    continue
  end if
  Compute centroid (x_s, y_s) and append to $\mathcal{C}$
end for
if $|\mathcal{C}| < 3$ **then**
  return `UNSAFE`
end if
Compute triangle $\mathcal{T}$ from: bottom-left: $\min_x$, bottom-right: $\max_x$, apex: $(\text{mean}_x, \min_y)$ of $\mathcal{C}$
Count $n_{\text{in}} \leftarrow$ number of points in $\mathcal{C}$ inside $\mathcal{T}$
if $n_{\text{in}} \geq 3$ **then**
  return `SAFE`
else
  return `UNSAFE`
end if

analyzes front-view images captured from the warehouse. The method makes use of the Segment Anything Model (SAM) for segmentation, followed by a centroid-based triangle inclusion test that determines whether bars are properly aligned with a predefined safe region (see Algorithm 2). The key assumption is that safely stacked bars should appear concentrated within a virtual support triangle, a geometrically defined region approximating the expected spatial distribution of correctly braced bar ends. If a sufficient number of bar-end centroids fall within this triangle, the configuration is classified safe.

Each front-view image is processed by a pre-trained SAM mask generator configured for high segmentation precision. The algorithm identifies candidate bar ends by filtering the generated masks based on their pixel area and circularity, properties that indicate compact and rounded support elements. After extracting valid bar centers from the image, the algorithm attempts to form a support triangle by selecting three reference points: the leftmost-bottom, rightmost-bottom, and topmost-center among the de-



(a) Top-view support masks

(b) Front-view safety triangle

Figure 3: Qualitative results of (a) dual-SAM support detection in top views (Green: the detected supports, Red: the not safe box, Blue: the safe box), and (b) triangle-based safety validation in front views.

tected bar coordinates. This triangle is then used as a geometric proxy for evaluating structural support. If three or more detected bar ends are found within the triangle, the bay is classified as structurally SAFE. Otherwise, it is marked as UNSAFE, indicating insufficient frontal bracing. Each image is visually annotated with this classification and saved for operator review. Finally, a CSV report summarizing per-image safety status and detection counts is generated to support large-scale batch analysis.

This frontal safety check complements the top-view analysis by enforcing a spatial constraint on bar placement. Together, the two modules form a comprehensive safety validation system, operating on dual views to ensure structural compliance.

5 OBTAINED RESULTS

This section presents both qualitative and quantitative evaluations of the proposed dual-SAM and triangle-based safety monitoring framework. The system has been tested on real-world video footage collected in operational steel bar storage facilities under varying environmental conditions. The results demonstrate the framework's ability to perform robust and interpretable safety validation from both top- and front-view perspectives.

5.1 Qualitative Evaluation

Top-View Support Detection. Figure 3a illustrates representative results of the top-view analysis. Green masks correspond to small-scale structural support elements detected via SAM, while red regions indicate initially unsafe bulk areas. Regions satisfying the proximity reclassification criteria, i.e., those in contact with at least three green masks, are re-annotated in blue to denote structural support.

Across multiple scenarios, the proposed dual-resolution segmentation approach consistently captured fine structural details (e.g., inclined supports and wedges), even under partial occlusion and non-uniform lighting. The use of morphological dilation and mask proximity significantly reduces false negatives, particularly in cluttered layouts. The resulting visual overlays offer a high degree of interpretability and enable clear identification of safety-critical zones for operator intervention or automated alerts.

Front-View Safety Triangle. Figure 3b presents examples of the triangle-based safety validation applied to front-view frames. Detected bar-end centroids are plotted as yellow points, while the computed support triangle is shown in cyan. Bays with three or more centroids located within the triangle are classified as "SAFE" (annotated in green), whereas those with insufficient frontal support are marked "UNSAFE" (annotated in red).

Table 1: Comparison of bar and triangle detection and safety classification.

Method	Avg. bars	Avg. triangles	Safety acc. (%)	Bar FP	Tri. FP	Time (ms)
SAM (Lin and Ferrari, 2024)	12.3	9.5	60	1.2	1.0	100
U-Net (Ronneberger et al., 2015)	11.8	6.9	50	0.9	1.3	50
Edge + Hough (Kälviäinen et al., 1995)	2.4	0.1	20	1.5	3.4	70

This method proved effective in distinguishing correctly stacked configurations from potentially hazardous ones. It was particularly robust in identifying over-extended bars or unevenly braced stacks, where traditional methods based solely on segmentation may fail to account for geometric safety constraints.

5.2 Quantitative Evaluation and Observations

To further assess the effectiveness of the proposed system, we performed a comparative evaluation involving three methods: (1) SAM - the one proposed in this work, (2) U-Net - a standard convolutional segmentation model, and (3) edge-based detection with probabilistic Hough transform - a classical geometric approach.

Each method was applied to a dataset of annotated top-view images, and their performance was measured across multiple safety-relevant metrics to assess their effectiveness in detecting both structural components (bars and supporting triangles) and their ability to correctly classify scenes as safe or unsafe.

The dataset includes manually annotated ground truth labels for the positions of steel bars and supporting triangles. These annotations serve as the basis for computing detection accuracy and false positive rates.

Table 1 summarizes the results in terms of average detections per frame, false positives, safety classification accuracy, and inference time. Each column in Table 1 reports specific aspects of the performance of the evaluated methods:

- Average Bars (Avg. Bars): The average number of correctly detected steel bars per frame, compared against the annotated ground truth. Higher values generally indicate better detection completeness.
- Average Triangles (Avg. Triangles): The average number of ground-truth support triangles correctly identified per frame. This metric reflects the method's ability to infer stable structural configurations, which are critical for safety assessment.
- Safety Accuracy (Safety Acc.) (%): The percentage of frames for which the method correctly classified the scene as either safe or unsafe based on the geometric reasoning applied to the detected structures. This is the final downstream task.
- Bar False Positives (Bar FP): The average number of bars detected per frame that do not correspond to any annotated ground truth bar. A lower value indicates higher precision.
- Triangle False Positives (Tri. FP): The average number of detected triangles per frame that are not supported by actual structural elements in the ground truth. High false positives can lead to erroneous safety classification (e.g., falsely declaring unsafe setups as safe).
- Time (ms): The average inference time per frame in milliseconds, including segmentation, post-processing, and safety classification. This gives an indication of the method's suitability for real-time applications.

We can see from the obtained results that: **SAM** achieved the highest accuracy in triangle detection, contributing to more reliable safety classification (60%, the average result among the tested methods). However, it incurred the highest computational cost. **U-Net** demonstrated a good balance between accuracy and efficiency, with moderate false positives and acceptable safety classification results (50%). **Edge + Hough** was significantly faster but suffered from low detection rates and poor classification accuracy (20%), likely due to its sensitivity to noise and lack of learned representations.

While the SAM-based approach showed promising structural detection capabilities, a safety classification accuracy of 60% indicates that substantial room for improvement remains. This result should be interpreted as a first-step baseline rather than a conclusive performance ceiling.

All methods demonstrated the capability to operate near real-time (10+ fps), but trade-offs between accuracy and performance must be considered for deployment in live monitoring systems.

These results highlight the trade-off between detection quality and computational cost. While edge-based methods offered lower latency, their limited precision and geometric inference capabilities rendered them unsuitable for reliable safety monitoring in realistic scenarios. In contrast, the SAM-based

model approach provides a balanced compromise between robustness, interpretability, and runtime efficiency, making it suitable for industrial deployment.

Next Steps: We acknowledge that the current evaluation is limited by dataset size and the scope of reported metrics. To strengthen the quantitative analysis, we plan to significantly expand the annotated dataset and compute standard detection metrics such as precision, recall, and F1-score for each stage (bar detection, triangle inference, safety classification). This broader evaluation will provide a more comprehensive understanding of each method's strengths and failure modes, and help guide future improvements in model architecture and rule design for industrial safety validation.

6 CONCLUSION

We proposed an annotation-light vision framework for real-time safety validation of steel bar storage in outdoor industrial environments. By combining dual-resolution zero-shot segmentation using SAM with lightweight geometric reasoning, the system assesses structural support from top and front views with no manual labeling.

Key contributions include: (i) multi-scale SAM mask generation for detecting both fine supports and bulk materials, (ii) morphological proximity rules for lateral support inference, (iii) triangle-based validation from frontal views, and (iv) efficient implementation suitable for real-world deployment.

Our method addresses key limitations of prior work by avoiding task-specific annotations, handling multi-scale structures, and offering interpretable, geometry-driven safety decisions. Experimental results on real warehouse footage show reliable performance under challenging conditions like occlusion and clutter.

Future work includes extending to more complex stacking scenarios, adding temporal smoothing, and integrating multi-camera fusion. We also plan to explore self-supervised fine-tuning of SAM for improved low-contrast performance. This work lays the foundation for fully automated structural safety monitoring in heavy-industry logistics.

ACKNOWLEDGEMENT

The COGNIMAN project[1], leading to this paper, has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101058477.

REFERENCES

Cen, J., Fang, J., and Shen, W. (2023). Segment anything in 3d with radiance fields. In *Proceedings of ICCV*.

Duda, R. and Hart, P. (1972). Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15.

Eiffert, S., Wendel, A., and Kirchner, N. (2021). Toolbox spotter: A computer vision system for real-world situational awareness in heavy industries. In *IEEE Conference on Automation Science and Engineering (CASE)*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of ICCV*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A., Lo, W.-Y., Dollár, P., and Girshick, R. (2023). Segment anything. In *Proceedings of ICCV*.

Kälviäinen, H., Hirvonen, P., Xu, L., and Oja, E. (1995). Probabilistic and non-probabilistic hough transforms: overview and comparisons. *Image and Vision Computing*, 13(4):239–252.

Lee, S. and Kim, H. (2021). Geometric primitive detection for structural support analysis. In *Proceedings of ICRA*.

Lin, X. and Ferrari, V. (2024). Sam-6d: Zero-shot 6d object pose estimation with segment anything. In *Proceedings of CVPR*.

Patel, R. and Gupta, S. (2020). Automated safety violation detection in manufacturing through vision ai. *IEEE Transactions on Industrial Informatics*, 17(5):3502–3512.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Smith, J. and Lee, P. (2019). Vision-based automation and safety in industrial environments: A survey. *IEEE Transactions on Automation Science and Engineering*, 16(4):1548–1565.

Wu, Y. and Zhang, X. (2019). Multi-scale image segmentation using deep learning for industrial applications. *Pattern Recognition Letters*, 120:109–116.

Zhang, L., Chen, Y., and Zhao, J. (2022). Proximity-based support verification in robotic assembly. In *Proceedings of IROS*.

[1] www.cogniman.eu