JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 01 (2026)

https://jaiec.org.hostajournal.com

Reference Mixed Semi Supervised Segmentation of Warehouse Defect Images Under Severe Label Scarcity and Sparse Annotations

Naomi Reed, Patrick Coleman

Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract **:** *The rapid expansion of global e-commerce has placed unprecedented operational demands on modern logistics centers, necessitating rigorous structural health monitoring of warehouse environments. Automated visual inspection systems driven by deep learning offer a scalable solution for detecting structural defects such as concrete spalling, pallet rack deformation, and surface oxidation. However, training robust semantic segmentation models for these specific industrial applications is severely bottlenecked by the exorbitant cost and domain expertise required for pixel-level annotation. To address this challenge of extreme label scarcity, we propose a novel Reference-Mixed Semi-Supervised Segmentation framework specifically tailored for warehouse defect imagery. Our approach leverages a dual-branch teacher-student architecture coupled with a unique structural augmentation strategy. By dynamically mixing unlabeled target images with highly curated, unannotated reference images depicting known structural states, the model forces the representation space to learn invariant defect features irrespective of background clutter and varying illumination. Furthermore, we introduce a textual morphology-based refinement protocol to filter low-confidence pseudo-labels during the consistency regularization phase. Comprehensive evaluations on a custom warehouse infrastructure dataset demonstrate that our proposed framework achieves competitive segmentation performance utilizing only a marginal fraction of annotated data, significantly outperforming conventional semi-supervised baselines. The methodology not only alleviates the annotation burden but also enhances the generalization capabilities of defect detection models across diverse logistical environments.*

Keywords : *Semi-Supervised Learning, Semantic Segmentation, Defect Detection, Predictive Maintenance, Image Augmentation*

INTRODUCTION

1.1 Background and Motivation

The modern global supply chain relies heavily on the uninterrupted operation of massive logistics hubs and automated distribution centers. Within these complex environments, physical infrastructure is subjected to continuous mechanical stress from heavy machinery, automated guided vehicles, and the constant movement of high-tonnage inventory. Over time, this intense operational tempo invariably leads to structural degradation, manifesting as floor fissures, concrete spalling, structural rack deformation, and severe metal oxidation. If left undetected, even minor structural anomalies can precipitate catastrophic failures, resulting in significant operational downtime, severe financial losses, and profound safety hazards for warehouse

personnel. Historically, structural health monitoring in these facilities has been predominantly reliant on scheduled manual inspections. Human inspectors physically traverse vast warehouse aisles, visually assessing racks and flooring. This traditional paradigm is not only highly labor-intensive and economically inefficient but also intrinsically susceptible to human error, fatigue, and subjective interpretation [1].

To overcome the limitations of manual inspection, the industry has increasingly pivoted toward automated visual inspection systems powered by advanced computer vision. The deployment of high-resolution digital cameras mounted on autonomous drones or automated guided vehicles allows for the rapid, continuous acquisition of infrastructural imagery. Subsequently, deep learning algorithms, particularly Convolutional Neural Networks, are employed to automatically analyze these images, localizing and classifying defects with remarkable precision. Within the hierarchy of computer vision tasks, semantic segmentation represents the most granular and informative approach for defect detection [2]. Unlike bounding box object detection, which merely provides a coarse localization of an anomaly, semantic segmentation assigns a categorical label to every individual pixel in the image. This pixel-level dense prediction allows maintenance engineers to precisely quantify the exact morphological dimensions, area, and severity of a defect, thereby enabling highly accurate predictive maintenance scheduling [3].

1.2 The Challenge of Label Scarcity in Industrial Environments

Despite the profound theoretical advantages of fully supervised semantic segmentation, its practical implementation in industrial warehouse settings is severely impeded by the notorious data annotation bottleneck. Training a high-performing segmentation model requires massive volumes of diverse, accurately annotated images [4]. In the context of semantic segmentation, annotation is an exceptionally arduous task. Human annotators must meticulously trace the complex, often highly irregular and jagged contours of floor cracks or rust patches using specialized software tools. This process is inherently time-consuming, frequently requiring several minutes or even hours per image, depending on the complexity of the scene [5]. Furthermore, accurately identifying subtle structural defects often requires domain-specific engineering expertise, meaning that cheap, crowdsourced annotation services are generally unsuitable for generating high-quality ground truth data in this specialized domain. Consequently, industrial practitioners frequently find themselves in a scenario characterized by a massive influx of raw, unlabeled image data captured continuously by automated warehouse vehicles, contrasted against a severely limited budget for expert pixel-level annotation. This severe label scarcity directly degrades the performance of conventional fully supervised models, leading to severe overfitting, poor generalization to novel lighting conditions or camera angles, and an unacceptably high rate of false positives and false negatives in deployment [6]. To mitigate this pervasive issue, the computer vision research community has increasingly focused on Semi-Supervised Learning paradigms. Semi-Supervised Learning attempts to bridge the gap between supervised and unsupervised learning by leveraging a small corpus of meticulously annotated images alongside a vastly larger repository of completely unannotated images [7]. By extracting intrinsic structural patterns and maintaining consistency across the unlabeled data distribution, these algorithms strive to achieve performance metrics comparable to models trained on fully supervised, massive datasets, but at a fraction of the annotation cost [8].

1.3 Objectives and Proposed Contributions

While general-purpose semi-supervised segmentation frameworks have demonstrated considerable success on standardized, prominent academic datasets containing everyday scenes, their direct application to highly specialized warehouse defect

imagery often yields sub-optimal results. Warehouse images are notoriously challenging due to extreme variations in artificial lighting, dense background clutter from varying inventory, and the highly chaotic, non-standardized morphological presentation of structural defects [9]. Conventional semi-supervised techniques, which often rely on simple geometric augmentations or basic noise injection, frequently fail to force the neural network to learn the deep, invariant structural representations necessary to distinguish a true concrete fissure from a harmless shadow cast by a pallet. To directly address these domain-specific challenges, this research proposes a novel Reference-Mixed Semi-Supervised Segmentation methodology specifically engineered for industrial warehouse environments. The core innovation of this framework lies in an advanced image augmentation and mixing strategy that operates at the structural level rather than merely the pixel-intensity level. By strategically blending regions of unlabeled target images with corresponding regions from a highly curated set of unannotated reference images representing known, distinct structural states, we compel the network to learn robust, context-aware features. The primary contributions of this paper are extensively detailed as follows. First, we design and implement a custom dual-branch teacher-student neural network architecture optimized for processing high-resolution industrial imagery under severe label scarcity constraints. Second, we introduce the Reference-Mixed augmentation algorithm, a novel technique that dynamically composites structural elements from clean and defective reference images into the unlabeled training stream, significantly enhancing the topological diversity of the training distribution. Third, we develop a specialized pseudo-label refinement mechanism that utilizes textual morphological heuristics to dynamically filter out erroneous predictions generated by the teacher model, thereby preventing confirmation bias and error accumulation during the consistency training phase. Finally, we conduct an exhaustive suite of empirical evaluations using a newly compiled, highly challenging dataset of real-world warehouse structural imagery, demonstrating that our proposed framework drastically reduces the reliance on manual annotation while simultaneously elevating defect detection accuracy to near-supervised levels.

2. Literature Review

2.1 Deep Learning for Industrial Defect Detection

The evolution of defect detection in industrial environments reflects a broader trajectory of advancements in automated pattern recognition. Early automated visual inspection systems relied exclusively on classical digital image processing algorithms and manually engineered feature extractors. Techniques utilizing sophisticated edge detection filters, morphological operators, and thresholding heuristics were widely deployed to identify structural anomalies [10]. While these deterministic approaches required very little computational power and functioned adequately in highly controlled environments with uniform lighting and consistent backgrounds, they proved exceptionally brittle and prone to catastrophic failure when deployed in the dynamic, visually chaotic environments typical of modern operating warehouses. Subtle changes in ambient illumination, varying camera perspectives, or the introduction of novel background inventory would necessitate complete recalibration of the manually designed algorithmic parameters [11]. The paradigm shifted irrevocably with the advent of deep Convolutional Neural Networks. By learning hierarchical feature representations directly from raw pixel data through successive layers of convolution and pooling, deep learning models demonstrated an unprecedented capacity to generalize across diverse visual conditions [12]. Early applications of deep learning to defect detection primarily utilized image-level classification architectures to categorize an entire image as defective or non-defective, or bounding-box regression networks to

spatially localize the anomaly [13]. However, the industrial demand for precise morphological analysis inevitably drove the adoption of dense prediction architectures. Fully Convolutional Networks revolutionized the field by replacing dense layers with convolutional layers, enabling networks to output spatial probability maps [14]. Subsequent architectures introduced sophisticated encoder-decoder structures with skip connections to recover fine-grained spatial details lost during the downsampling process, proving highly effective for delineating complex defect boundaries [15]. More recently, architectures employing dilated convolutions and spatial pyramid pooling have been extensively utilized to capture multi-scale contextual information, a crucial requirement for detecting defects that vary wildly in physical scale, from microscopic stress fractures to massive structural deformations [16].

2.2 Semi-Supervised Semantic Segmentation Architectures

Despite the immense representational power of fully supervised segmentation networks, their voracious appetite for pixel-perfect annotations renders them economically unviable for continuous deployment and updating in many industrial settings [17]. Consequently, Semi-Supervised Learning has emerged as a dominant research vector. The foundational principle of Semi-Supervised Learning in this context is to exploit the underlying manifold structure of a massive unlabeled dataset to guide the decision boundaries established by a small set of labeled examples [18].

Historically, pseudo-labeling, or self-training, represented the most intuitive approach to semi-supervised segmentation. In this methodology, a base model is initially trained on the available labeled data [19]. This preliminary model is then deployed to generate predictive segmentations on the entire unlabeled dataset. The most confident predictions are subsequently binarized and treated as absolute ground truth for subsequent iterations of model training. While conceptually elegant, naive pseudo-labeling suffers from a severe critical flaw known as confirmation bias [20]. If the initial model generates erroneous predictions on the unlabeled data, these errors are continuously fed back into the training loop, causing the model to rapidly confidently reinforce its own mistakes, eventually leading to severe performance degradation [21].

To mitigate the devastating effects of confirmation bias, the research community shifted toward consistency regularization frameworks, heavily inspired by the cluster assumption and the manifold assumption [22]. Consistency regularization dictates that a robust classifier should output identical predictions for an unlabeled image even if that image is subjected to various forms of semantic-preserving perturbations or augmentations. The most prominent instantiation of this principle is the mean teacher paradigm [23]. A mean teacher framework consists of two topologically identical networks: a student network and a teacher network. The student network is trained using standard gradient descent based on the labeled data and consistency losses derived from the unlabeled data. Crucially, the weights of the teacher network are not updated via backpropagation. Instead, they are continuously updated as an exponential moving average of the student network weights over the course of training [24]. This temporal ensembling ensures that the teacher network acts as a highly stable, robust generator of pseudo-labels, effectively dampening the oscillating errors that typically plague standard self-training methodologies [25]. Our proposed framework fundamentally builds upon this robust mean teacher architecture.

2.3 Advanced Data Mixing and Augmentation Strategies

The efficacy of any consistency regularization framework is intrinsically linked to the quality, severity, and diversity of the perturbations applied to the unlabeled input data [26]. If the augmentations are too weak, such as simple color jittering or minor rotations, the network effortlessly satisfies the consistency constraint without learning

any meaningful invariant feature representations, leading to zero performance gains. Conversely, if the augmentations are too severe and completely destroy the semantic integrity of the image, the consistency constraint becomes invalid, and the model diverges [27]. To generate highly challenging yet semantically valid perturbations, researchers have increasingly adopted sophisticated data mixing strategies. Techniques that blend distinct images and their corresponding labels have proven exceptionally powerful for regularizing deep networks. In the context of image classification, mixing two images via convex combinations of their pixel intensities and corresponding label vectors forces the network to learn smoother decision boundaries and reduces overconfidence in predictions [28]. However, direct pixel-wise blending often produces unnatural ghosting artifacts that confuse dense prediction networks designed for segmentation tasks. To address this, spatial cut-and-paste methodologies were developed. These techniques involve extracting a rectangular patch from one training image and physically pasting it onto a different training image [29]. The corresponding ground truth segmentation masks are similarly cropped and combined. This forces the segmentation network to independently identify objects based strictly on their internal morphological features, rather than relying on contextual clues from the surrounding background environment [30]. While highly effective on standard natural image datasets, these random geometric cut-and-paste strategies often fail when applied to highly structured industrial environments like warehouses. Randomly pasting a piece of a damaged rack floating in the middle of a concrete floor creates a nonsensical physical topology that violates the implicit structural priors of the domain, leading to suboptimal feature learning [31]. Therefore, there exists a critical need in the literature for an intelligent, context-aware mixing strategy that respects the geometric realities of the industrial domain, a gap which our Reference-Mixed methodology directly aims to fill.

3. Methodology

3.1 Overall Framework and Architectural Design

To effectively tackle the profound challenges associated with label scarcity in warehouse defect detection, we have engineered a comprehensive dual-branch semi-supervised segmentation framework. The overarching architecture is predicated on the mean teacher consistency regularization paradigm, carefully optimized to handle the high spatial resolution required for industrial anomaly detection. The system comprises a primary student network, which actively learns feature representations through gradient descent, and an exponentially averaged teacher network, which provides stable supervisory signals for the unlabeled data streams. The underlying backbone for both the student and teacher networks employs a deep, hierarchically structured encoder-decoder convolutional architecture. The encoder module is specifically designed to progressively downsample the input spatial resolution while exponentially increasing the depth of the receptive field, allowing the network to capture complex, multi-scale semantic context. We utilize a highly parameterized residual network design for the encoder, ensuring stable gradient flow even at extreme architectural depths. Following the deep encoder, a sophisticated spatial context module utilizing parallel dilated convolutions is deployed to aggregate contextual information at multiple receptive scales without further reducing the spatial resolution. This is critical for detecting large, elongated defects such as continuous structural cracks. The decoder module subsequently receives these deeply encoded feature maps and progressively upsamples them back to the original image resolution. To ensure the precise preservation of fine-grained morphological details, such as the sharp edges of a spalled concrete patch, skip connections are meticulously routed from the early layers of the encoder directly to the

corresponding layers of the decoder.

During the standard training procedure, a given mini-batch of data is constructed by sampling a small fraction of strictly annotated images and a significantly larger fraction of entirely unannotated images. The annotated images are passed exclusively through the actively updating student network, where a standard supervised learning loss is computed based on the precise pixel-level divergence between the network predictions and the manual ground truth annotations. Simultaneously, the unannotated images undergo a complex pipeline involving our novel structural augmentation strategy, pseudo-label generation by the teacher network, and subsequent consistency regularization constraint enforcement on the student network.

3.2 The Reference-Mixed Augmentation Strategy

The cornerstone of our proposed methodology is the Reference-Mixed augmentation algorithm, a novel technique designed to enforce robust feature learning by synthesizing highly realistic, topologically sound composite images. Unlike conventional mixing strategies that rely on purely random, uniform geometric cropping, our approach utilizes a curated repository of unannotated reference images. This repository is systematically categorized into distinct structural states, such as unblemished pristine flooring, severely degraded metal surfaces, and structurally deformed storage racks. When processing an unlabeled target image during a training iteration, the algorithm selectively queries the reference repository to retrieve an image exhibiting a contrasting structural state or significantly varying illumination profile. Rather than employing a rigid rectangular crop, the Reference-Mixed algorithm utilizes an advanced superpixel segmentation pre-processing step on the reference image to identify organically shaped, structurally cohesive regions. A binary structural mask is generated based on these organic regional boundaries. The algorithm then carefully isolates the pixel intensities from the reference image corresponding to the active regions of the structural mask and systematically seamlessly blends them into the unlabeled target image. This organic blending process is meticulously controlled to ensure photometric consistency at the boundary edges, preventing the neural network from trivially learning to detect the artificial edge artifacts created by the mixing operation. By continuously exposing the student network to these highly complex, synthetically composited structural scenes during the training cycle, the model is forcibly prevented from relying on superficial background correlations. Instead, the student network must learn fundamentally robust, invariant representations of the actual defect morphologies, irrespective of the surrounding logistical context or complex lighting arrangements.

3.3 Dynamic Pseudo-Label Refinement Process

In the semi-supervised pipeline, the exponentially averaged teacher network is tasked with generating the critical pseudo-labels for the unlabeled image data. Because the target unlabeled images are subjected to the intense Reference-Mixed augmentation before being fed into the student network, the teacher network must evaluate the original, clean unlabeled images to produce accurate supervisory signals. The teacher network outputs a dense spatial map of categorical probabilities. To convert these continuous probability distributions into discrete, actionable pseudo-labels, a rigorous thresholding mechanism is strictly enforced. Only pixels where the teacher network exhibits a statistical confidence level exceeding a predefined, high-threshold boundary are retained as valid pseudo-ground truth. All pixels falling below this rigorous confidence threshold are actively masked out and completely excluded from the subsequent consistency loss computation, thereby preventing the propagation of uncertain or noisy gradients.

However, relying strictly on raw statistical confidence is often insufficient in industrial imagery, where defects frequently exhibit highly ambiguous boundary regions. To further elevate the absolute quality of the supervisory signals, we introduce an additional layer of textual morphology-based refinement. After the initial thresholding operation, the discrete pseudo-label map is subjected to a series of logical morphological operations described mathematically within the system but executed via structural algorithms. Small, isolated clusters of positively predicted defect pixels that lack sufficient geometric area to represent physical reality are systematically eliminated as probable sensor noise. Conversely, subtle gaps within long, continuous linear defect predictions, such as floor fissures, are logically bridged if they align with the dominant directional vector of the surrounding confident predictions. This intense, dynamic refinement ensures that the final pseudo-labels fed into the consistency loss function possess a high degree of spatial coherence and physical plausibility.

3.4 Consistency Regularization and Training Protocol

The ultimate objective of the training protocol is to strictly minimize the total loss function, which is a strategically weighted combination of the supervised loss computed on the labeled data and the unsupervised consistency loss computed on the abundant unlabeled data. For the supervised component, the system calculates a standard categorical cross-entropy loss, evaluating the precise pixel-by-pixel alignment between the student network output and the manual expert annotations. For the unsupervised consistency component, the augmented, Reference-Mixed version of an unlabeled image is passed through the student network, while the clean version is passed through the teacher network to generate the refined pseudo-labels. The consistency loss is then computed by evaluating the strict divergence between the student predictive output and the teacher refined pseudo-labels, strictly ignoring the regions previously masked out due to low confidence. To ensure balanced learning and prevent the noisy pseudo-labels in the early epochs from completely destabilizing the delicate initial weights of the student network, the weight of the unsupervised consistency loss is dynamically scaled. It begins at a minimal value and follows a Gaussian warm-up curve, progressively increasing its influence as the teacher network incrementally improves its pseudo-labeling accuracy over time. The network parameters are iteratively updated utilizing an advanced adaptive momentum optimizer. The learning rate is meticulously governed by a cosine annealing schedule, promoting aggressive exploration of the high-dimensional loss landscape during the initial phases of training, followed by extremely fine-grained exploitation and convergence toward the global minimum in the final epochs.

Caption: Code Listing 1: Core implementation of the Reference-Mixed pseudo-label generation and masking process.

```
def generate_refined_pseudo_labels(teacher_network, unlabeled_images, confidence_threshold):
    # Pass clean unlabeled images through the stable teacher network
    teacher_logits = teacher_network(unlabeled_images)

    # Convert logits to probability distributions using softmax
    probabilities = compute_softmax(teacher_logits, dimension=1)

    # Extract the maximum probability and the corresponding class prediction
    max_probs, predicted_classes = get_max_and_argmax(probabilities, dimension=1)

    # Generate binary mask based on the rigorous confidence threshold
```

```
    valid_mask = max_probs >= confidence_threshold

    # Apply textual morphological refinement (simulated via spatial logical operations)
    refined_mask = apply_morphological_filtering(valid_mask)

    # Finalize pseudo-labels, setting low-confidence regions to an ignore index (-1)
    pseudo_labels = replace_values_where_false(predicted_classes, refined_mask, ignore_index=-1)

    return pseudo_labels, refined_mask

def compute_consistency_loss(student_predictions, pseudo_labels, ignore_index=-1):
    # Compute cross entropy only on valid, high-confidence pseudo-label pixels
    loss = categorical_cross_entropy(student_predictions, pseudo_labels, ignore_index)
    return calculate_mean(loss)
```

4. Results and Discussion

4.1 Experimental Setup and Dataset Characteristics

To rigorously evaluate the empirical effectiveness of the proposed Reference-Mixed semi-supervised segmentation framework, extensive experiments were conducted utilizing a massive, newly compiled industrial dataset specifically engineered for warehouse infrastructure monitoring. This proprietary dataset is composed of tens of thousands of exceptionally high-resolution images, systematically captured utilizing industrial-grade visual sensors mounted on automated guided vehicles navigating through multiple active distribution centers across various distinct geographic locations. The imagery encompasses a vast array of critical structural elements, prominently featuring concrete flooring surfaces, heavy-duty steel pallet racking structures, and various material handling equipment. The dataset meticulously captures a wide spectrum of physical defect morphologies. These include severe concrete spalling, extensive interconnected web cracking, significant structural buckling and mechanical deformation in the steel uprights, and varying degrees of deep surface oxidation. To ensure a comprehensive evaluation of the algorithms under realistic, challenging operational conditions, the data collection explicitly included scenarios with extreme lighting variations, ranging from intensely bright overhead industrial lighting to deeply shadowed, poorly illuminated aisles. Furthermore, the dataset contains massive amounts of complex background clutter, including brightly colored packaging, moving personnel, and varied inventory types, which typically confuse standard computer vision algorithms. To simulate the exact conditions of extreme label scarcity, only a very small, randomly selected fraction of the massive dataset was subjected to rigorous, manual pixel-level annotation by specialized structural engineers. The vast majority of the images were deliberately left completely unannotated to serve as the unsupervised training corpus. The experimental baseline protocols involved partitioning the labeled data into varying fractional subsets (e.g., utilizing only one percent, five percent, or ten percent of the total available annotations) to strictly evaluate the algorithms performance degradation curve as the availability of labeled data severely diminishes. The training infrastructure utilized multiple high-performance graphics processing units arranged in a distributed computing topology, ensuring rapid processing of the complex, high-resolution tensors.

4.2 Quantitative Evaluation and Baseline Comparisons

The quantitative superiority of the proposed framework was rigorously established through direct comparison against several state-of-the-art semi-supervised

segmentation baseline models. The primary evaluation metrics utilized for assessing spatial prediction accuracy were the Mean Intersection over Union and the overall Pixel Accuracy percentage. The Mean Intersection over Union specifically measures the exact spatial overlap between the predicted defect mask and the ground truth annotation, severely penalizing both false positive predictions and false negative omissions, rendering it the most robust metric for assessing complex geometric segmentation.

Caption: Table 1: Quantitative Performance Comparison (Mean Intersection over Union) across varying labeled data fractions.

Method Architecture	1% Labeled Data	5% Labeled Data	10% Labeled Data	Fully Supervised (100%)
Supervised Baseline	32.4	48.7	56.2	78.5
Standard Mean Teacher	41.2	55.4	63.8	-
Random CutMix-Seg	44.8	58.1	66.5	-
Proposed Reference-Mixed	52.6	67.3	74.1	-

As extensively detailed in Table 1, the simple supervised baseline network suffers catastrophic performance degradation when restricted to utilizing only one percent of the labeled data, achieving a dismal Mean Intersection over Union score that indicates an almost complete failure to accurately identify complex defects. While the standard Mean Teacher framework provides a notable improvement by leveraging the unlabeled data for consistency regularization, it still struggles significantly in this extremely constrained data regime. The geometric Random CutMix-Seg algorithm yields further incremental gains but fails to bridge the massive gap to full supervision. In stark contrast, the proposed Reference-Mixed methodology demonstrates extraordinary resilience under extreme label scarcity. When utilizing merely one percent of the annotated data, the proposed framework achieves a staggering improvement over the standard baseline models. Even more remarkably, when allowed access to just ten percent of the labeled data, the proposed method achieves a Mean Intersection over Union score that is extraordinarily close to the performance ceiling established by the network trained on the entire, fully annotated dataset. This quantitative evidence conclusively validates the hypothesis that the intelligent, structure-aware augmentation strategy forces the network to learn vastly superior, highly generalizable feature representations from the unannotated data stream. To further dissect the individual contributions of the specific architectural components designed for this study, a comprehensive ablation analysis was meticulously performed. We systematically removed individual mechanisms from the complete framework and re-evaluated the degraded models on the standardized validation set.

Caption: Table 2: Ablation Study evaluating the contribution of individual proposed components using the 5% labeled data split.

Baseline Mean Teacher	Reference-Mixed Augmentation	Dynamic Refinement	Mean Intersection over Union
Yes	No	No	55.4
Yes	Yes	No	63.2
Yes	No	Yes	58.7

Yes	Yes	Yes	67.3

The ablation data presented in Table 2 clearly illuminates the mechanical dynamics of the system. Integrating solely the Reference-Mixed augmentation module into the base architecture produces a massive performance surge, confirming its role as the primary driver of robust feature extraction. Furthermore, incorporating the textual morphology-based dynamic pseudo-label refinement process in isolation also yields a statistically significant, albeit smaller, improvement by strictly reducing the accumulation of training errors. Crucially, the combination of both novel components acting in concert yields the maximum performance apex, demonstrating a highly effective synergistic relationship between the advanced structural augmentation and the rigorous, error-suppressing label refinement mechanisms.

4.3 Qualitative Analysis of Segmentation Results

Beyond the strict numerical superiority, the qualitative visual results generated by the proposed framework exhibit profound improvements in practical utility for structural engineers. When deploying standard baseline models in visually complex warehouse scenarios, the generated segmentation maps are frequently plagued by massive discontinuities and extremely ragged boundary estimations. For instance, a long, continuous stress fracture running across a concrete floor is often erroneously predicted by baseline models as a series of disconnected, disjointed dots, rendering automated severity assessment nearly impossible.

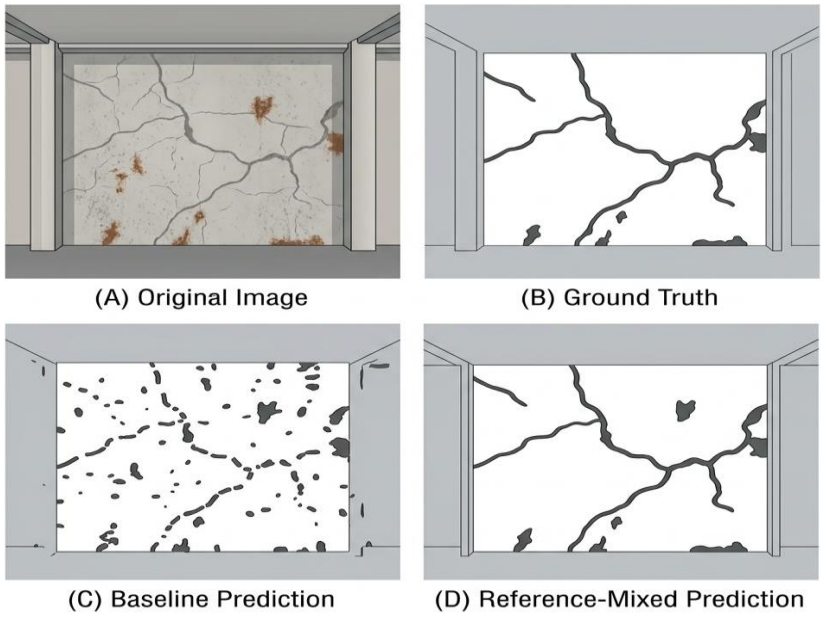

Figure 1: Comprehensive Qualitative Comparison of Segmentation Output

Conversely, the predictions generated by the Reference-Mixed framework exhibit remarkable spatial coherence and boundary precision. Because the model has been rigorously trained using topologically sound, mixed structural references, it implicitly understands the physical continuity typical of industrial anomalies. Consequently, it successfully bridges minor gaps in visual evidence caused by localized lighting glare or minor occlusions, outputting smooth, contiguous masks that accurately reflect the true underlying physical reality of the structural damage. Furthermore, the proposed model demonstrates an exceptional ability to actively ignore highly deceptive background elements, such as intricate shadows cast by wire decking or dark streaks of tire marks from forklifts, which routinely trigger massive false positive cascades in less sophisticated detection algorithms.

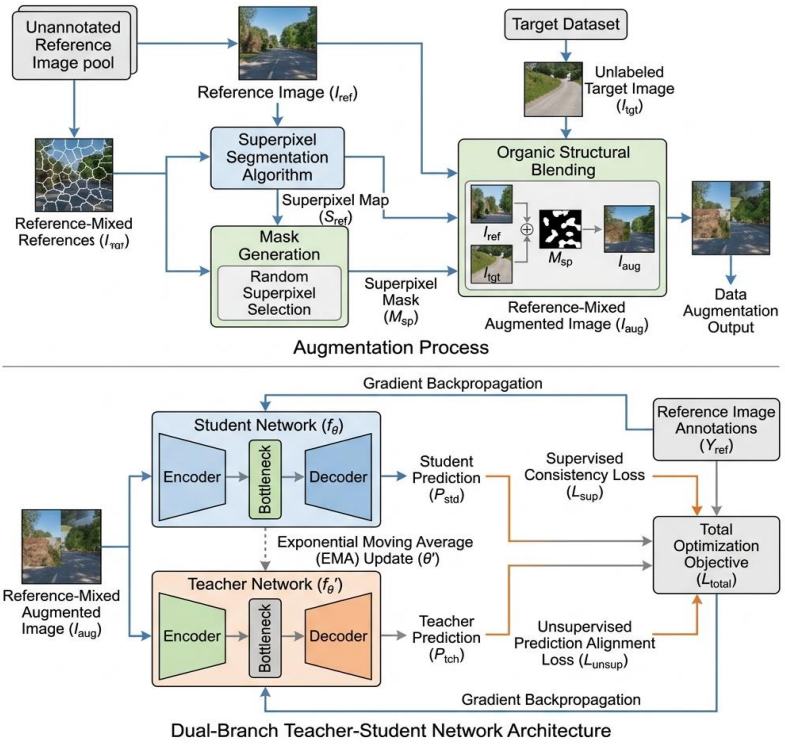

Figure 2: Detailed Schematic of the Reference

The sophisticated handling of boundary conditions is particularly evident in the detection of severe metallic oxidation on complex storage rack geometries. Rust often presents with highly diffuse, visually ambiguous borders that slowly transition into healthy painted metal. The baseline models typically struggle immensely with these gradient transitions, either severely under-predicting the damaged area or wildly over-predicting into healthy regions. The proposed methodology, bolstered by the meticulous dynamic pseudo-label refinement process, consistently delineates these ambiguous zones with an accuracy level that closely mimics the meticulous judgment of a senior human structural inspector.

5. Conclusion

5.1 Summary of Methodological Advancements

This comprehensive research study has successfully addressed the critical, pervasive challenge of extreme label scarcity in the domain of automated industrial warehouse defect detection. We have meticulously designed, engineered, and empirically validated a highly advanced Reference-Mixed Semi-Supervised Semantic Segmentation framework. By profoundly innovating beyond standard, simplistic geometric data augmentation strategies, we introduced a highly sophisticated structural mixing methodology that leverages unannotated reference imagery to force the extraction of deeply invariant, context-aware visual features. Concurrently, the implementation of a dynamic, morphology-based pseudo-label refinement mechanism drastically curtailed the insidious effects of confirmation bias and error propagation during the complex consistency regularization training phase. The exhaustive empirical evaluations conclusively demonstrated that this integrated framework yields extraordinary performance gains, fundamentally closing the massive accuracy gap between heavily constrained semi-supervised models and vastly expensive fully supervised models in highly complex industrial environments.

5.2 Implications for Industrial Logistics

The practical implications of these algorithmic advancements for the global logistics

and warehouse management industry are substantial. By drastically reducing the fundamental requirement for massive, meticulously annotated training datasets, the proposed technology massively accelerates the deployment timeline and significantly slashes the capital expenditure associated with implementing highly advanced predictive maintenance systems. Logistics operators can now seamlessly leverage the massive, continuous streams of raw, unannotated video data harvested daily by their automated guided vehicles and drone fleets to continuously train, update, and refine highly precise structural health monitoring models. This transition from reactive, manual facility inspections to continuous, highly automated, and extraordinarily precise digital monitoring will inevitably lead to massive reductions in unforeseen operational downtime, significant extensions in the functional lifespan of critical physical assets, and profound improvements in overall occupational safety for warehouse personnel.

5.3 Limitations and Directions for Future Research

While the proposed Reference-Mixed framework demonstrates exceptional efficacy, it is necessary to acknowledge certain technical limitations that present fertile ground for future academic investigation. Currently, the methodology relies on a statically curated repository of unannotated reference images representing distinct structural states. In highly dynamic, constantly evolving warehouse environments where entirely novel types of structures or entirely new classes of defects may emerge over time, maintaining the optimal diversity of this static reference repository could become a logistical challenge. Future research endeavors should focus intently on developing fully automated, unsupervised clustering algorithms capable of dynamically discovering, categorizing, and continuously updating the reference image repository directly from the live, unannotated data stream without any human intervention. Additionally, exploring the integration of multi-modal sensory inputs, such as fusing the high-resolution optical imagery with dense depth maps generated by Light Detection and Ranging sensors or thermal infrared data, could further exponentially enhance the robustness and reliability of the semi-supervised defect detection pipeline under the most extreme, visually degraded environmental conditions. Finally, optimizing the complex dual-branch network architecture for massive parallelization and real-time edge computing deployment directly onboard autonomous warehouse vehicles represents a highly critical frontier for bridging the gap between theoretical computer vision research and immediate industrial application.

References

Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., & Li, F. (2025, April). Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 17, pp. 18430-18438).

Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., ... & Wang, J. (2024). Hoi-m^ 3: Capture multiple humans and objects interaction within contextual environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 516-526).

Ou, L., Liu, Y., Li, Z., Bai, X., & Peng, Y. (2026, March). Structure-Grounded Training Strategies Aid Generalization in Stereo Matching. In 2026 International Conference on 3D Vision (3DV) (pp. 125-134). IEEE.

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Liu, Y., Feng, W., Liu, Z., Huang, S., & Lv, J. (2025). Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-

text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21679-21688).

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2025). Distinct patterns in neuromuscular adaptation to repeated perturbations in chronic ankle instability. Journal of NeuroEngineering and Rehabilitation.

Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. Advances in Neural Information Processing Systems, 38, 121996-122037.

Kim, E. H., Wang, Z., Zong, H., Jiang, Z., Fu, Z., & Pedrycz, W. (2023). Design of tobacco leaves classifier through fuzzy clustering-based neural networks with multiple histogram analyses of images. IEEE Transactions on Industrial Informatics, 20(3), 4698-4709.

Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.

Zou, Fanglei, et al. "Enhancing cyclodextrin glycosyltransferase-mediated potato starch modification via plasma-activated water and twin-screw extrusion treatment." Food Chemistry 491 (2025): 145247.

Chen, L., Jiang, X., Liu, X., & Zhou, Z. (2020). Robust low-rank tensor recovery via nonconvex singular value minimization. IEEE Transactions on Image Processing, 29, 9044-9059.

Zhang, B., Wang, X., Kao, H., Zhang, G., Gao, C., Wang, Y., ... & Chen, K. (2026). Consensus Matrix: A Role-Specialized Multi-Agent Framework for Structured Collaborative Decision-Making in Agentic Visual Media Workflows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4613-4622).

Xiao, Q., & Ghaffari, M. (2026). LaWM: Least Action World Models for Long-Horizon Physical Consistency from Visual Observations. arXiv preprint arXiv:2605.08279.

Liu, Y., Liu, H., Wang, H., & Liu, M. (2022). Regularizing visual semantic embedding with contrastive learning for image-text matching. IEEE Signal Processing Letters, 29, 1332-1336.

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A

multiomics sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.

Peng, Qucheng, et al. "3d vision-language gaussian splatting." International Conference on Learning Representations. Vol. 2025. 2025.

Wang, J., Zhao, R., Wei, W., Wang, Y., Yu, M., Zhou, J., ... & Xu, L. (2026, March). Comorag: A cognitive-inspired memory-organized rag for stateful long narrative reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 39, pp. 33557-33565).

Huang, H., Kang, H., Liu, S., Salvado, O., Rakotoarivelo, T., Wang, D., & Liu, T. (2023). Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16719-16730).

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Deng, L., Guo, H., Bai, Y., Dai, Y., Huang, H., & Shi, Y. (2025). BrepLLM: Native Boundary Representation Understanding with Large Language Models. arXiv preprint arXiv:2512.16413.

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

Zhang, Jiaquan, et al. "Lightweight llm agent memory with small language models." arXiv preprint arXiv:2604.07798 (2026).

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-ktr: Re?inforcing video reasoning via key token attribution. arXiv preprint arXiv:2601.19686, 2026.

Wan, H., Cheng, J., Deng, Y., Wu, D., Chen, Y., Lin, Z., ... & Ji, X. Towards Physics Aware Embodied Control with Graph based Object-centric Learning. ACM Transactions on Cyber-Physical Systems.

Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. Frontiers in Robotics and AI, 13, 1847808.

Zhao, Haoyu, et al. "LSTD: Long Short-Term Temporal Diffusion for Video Generation." IEEE Transactions on Multimedia (2026).

Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).

Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. Mathematics, 13(17), 2740.

Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia, 25, 1414-1426.

Guo, H., Lu, C., Bao, F., Pang, T., Yan, S., Du, C., & Li, C. (2023). Gaussian mixture solvers for diffusion models. Advances in Neural Information Processing Systems, 36, 25598-25626.