Journal of Applied Automation Technologies

Research Paper

Robust Visual Product Matching in Intelligent Warehouses via Siamese Networks and Adaptive Data Augmentation

Mohamed Al Khalifa[1] and Aziza Al Romaithi[1, *]

[1] School of Engineering, American University of Ras Al Khaimah, AURAK, Ras Al Khaimah, 10021, United Arab Emirates
*Corresponding author: a.romaithi@aurak.ac.ae

Abstract. In order to meet the ever-changing demands and higher accuracy requirements of intelligent warehouse management systems, automatic product identification technology will be added. The lack of stable visual matching in large-scale logistics is the main issue this paper aims to address. Lighting, occlusion, or any changes to the product can reduce the stability of the solution. Propose an enhanced Siamese network structure while establishing an augmented data pipeline to simulate warehouse changes. Strict labeling and partitioning rules have already been adopted, and a large number of real warehouse images have been used for system validation. In conditions of poor lighting or partial occlusion, the average exceeds 91%, with inference speeds below 50 milliseconds on typical hardware. Category-level accuracy and overall robustness have significantly improved compared to traditional convolutional and metric learning baselines. Adaptive decision logic and context-aware network training can reduce failure rates and ensure stable operation of visually similar or label-ambiguous products. The aforementioned findings lay the foundation for a highly scalable platform for the next generation of warehouse automation. To scale, continuous learning and various data types will be added in the future.

Keywords: *Visual Recognition, Siamese Network, Warehouse Automation, Data Augmentation, Product Identification, Robust Matching*

Received on 16 November 2025, Accepted on 10 March 2026, Published on 26 March 2026

Introduction

With the continuous development of automation technology, smart warehouses have been introduced to enhance the flexibility and resilience of the entire supply chain system [1]. Intelligent automated warehouses use computer vision and machine learning technologies to achieve various functions, including product recognition, automatic sorting, and real-time tracking of goods in transit [2]. As the scale and complexity of warehouses increase, the variety and quantity of products handled daily also increase significantly. Reliable and effective solutions are needed to accurately identify and match large datasets [3]. Vision-based systems have achieved good results in feature extraction and similarity computation through deep learning in controlled environments [4]. This field is striving to adapt to more realistic deployment environments; for example, many operational factors can reduce the performance of even the most advanced image-based recognition pipelines [5].

In the field of automated visual product matching, there are significant issues. The actual warehouse environment is usually not ideal; fluctuations in light intensity, severe occlusion caused by stacked goods, cluttered backgrounds, and products that look very similar are some of the issues [6]. Due to the stacking of products and other movements, the camera angle will change, making recognition more difficult [7]. The accuracy or recognition error rate is relatively low, and even minor modifications to the packaging design or label can lead to this situation [8]. Early methods based on manual features were often unreliable, but newer deep learning models are relatively sensitive to the rapid changes in logistics inventory distribution and data scarcity [9]. There are many good methods for data augmentation, but sometimes they cannot generate enough diversity in visual data to support the model's operation in large-scale dynamic warehouses [10].

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

This paper proposes a Siamese network framework specifically designed for visual matching of automated warehouse products. Through complex data augmentation and deep metric learning, a new method is introduced to enhance the model's robustness to different environments and products. The framework demonstrates high efficiency and accuracy in complex warehouse environments. Propose a feasible method to build a large-scale deployable visual matching system.

Related Technologies

Machine Vision in Logistics Management

In order to meet the growing demand for large-scale, high-efficiency automation in the supply chain, machine vision technology in warehouse logistics is rapidly developing [11]. Modern warehouses use vision inspection systems to track and identify large-scale inventory networks in real-time. Machine vision was initially used to replace traditional barcode scanning with automatic cameras and fixed imaging stations to improve throughput, speed, and save labor. Due to the chaotic environment, the first attempt failed. The display and visibility of the product need to be very high. The increase in supply chain complexity can exacerbate the aforementioned issues, such as partially hidden goods, inconsistent labeling, or overlapping packaging materials [12].

Most of the aforementioned image processing techniques are used to identify products in logistics. The visual inspection for automated shelf restocking and shipment verification has already adopted edge detection, spot extraction, template matching, and simple histogram-based classification algorithms [13]. These methods are fundamentally unable to cope with environmental changes, such as fluctuations in lighting and position, as well as various cluttered backgrounds. Due to the frequent and unstable changes in the warehouse environment, the image quality is inconsistent, and basic preprocessing generates a large number of artifacts, which are difficult to identify or correct. So far, the scalability of traditional vision pipelines has not been achieved, and large and diverse warehouses will be unmanageable [14]. Advanced visual intelligence research in the logistics field is ongoing, and it is necessary to introduce more learning-capable frameworks to address the aforementioned issues [15].

Deep Metric Learning

Deep metric learning is widely used in computer vision recognition tasks to improve the efficiency of distinguishing functions such as fine-grained visual classification [16]. Map the samples to the learned feature space. These distances between representations can be considered as visual or semantic distances. Good metrics are used for automatic product matching to find items that look similar, even if some parts are not visible or the packaging is different. Metric learning only considers the distance between samples, making it more suitable for open warehouse environments where new products can be added. In most cases, the original classifier is not optimized for these.

The parallel weight-sharing subnetwork architecture is very popular in many designs of Siamese networks, as it can independently extract features from paired images [17]. During the training phase, similarity measures such as Euclidean distance and cosine similarity are used to process the generated embeddings. Contrastive loss or triplet loss brings the encodings of the same items closer together while pushing the encodings of different items further apart. Unlabeled methods can be extended to unseen items and are very flexible in the expansion of the inventory pool. Difficult negative sample mining and adaptive boundary methods focus on complex instance pairs and discriminative boundaries to improve efficiency [18]. In order to enhance the ability to represent noise in real-world images, recent advancements include multi-scale feature extraction, channel attention mechanisms, and ensemble models.

In practice, logistics also encountered issues with training programs and data engineering, and the main architecture did not directly address these problems [19]. Data augmentation techniques have been used to improve the robustness of the metric space to lighting changes and unusual occlusions (such as color jitter, perspective distortion, and domain randomization overlay). Sample diversity is too low, or informative samples are not carefully selected when using triplet loss, which can lead to instability due to data distribution [20]. One of the main reasons for the great success of deep metric learning in warehouse automation is the cross-research results based on network design, sampling techniques, and data strategies [21].

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

Previous Approaches to Product Matching

The visual product matching in logistics has transitioned from traditional manual model construction to comprehensive deep learning pipelines [22]. Early systems achieved automatic detection through geometric descriptors, color histograms, edge matching, and local features similar to SIFT and SURF. Applicable in controlled environments, but due to the large number of SKUs, minor changes in packaging, and other frequent changes in global warehouse operations, especially in highly competitive environments, it cannot be used here. When logistics centers begin handling fast-moving consumer goods, this difference becomes more pronounced [23].

In the era of deep learning, convolutional neural networks (CNNs) are considered to improve the generalization performance of product recognition systems [24]. The aforementioned architecture improves the robustness of the product environment while reducing the need for manual feature engineering. Due to the limited number of old categories in the classification network, it is necessary to manually re-label and retrain any new packaging or SKU additions. To improve adaptation speed and reduce the amount of data labeling, methods based on metric learning classification networks, transfer learning, and few-shot learning are being researched. Open set and continual learning techniques are being used to adapt recognition systems to handle previously unseen new categories in real-world warehouses [25].

There are still some issues at present. Traditional and new recognition systems still face issues of data imbalance and rare categories, as warehouse images are affected by natural changes and are difficult to predict. Expanding the training set through synthetic data or active learning pipelines is feasible, but it may also introduce new errors and issues of overfitting or transferability. Efficiently and flexibly connecting visual product matching with picking, sorting, and inventory management on a large scale requires robust end-to-end integration with warehouse robots, which necessitates high-precision algorithms. The basic vision system must adapt to changes in the logistics environment to support efficiency improvements. These systems must provide reliable, flexible, and universally applicable product matching capabilities. To address the current issues, collaboration between academia and the business sector still needs to be strengthened.

Methodology

Enhanced Siamese Network Structure

In the automatic visual product matching embedding architecture in smart warehouses, attention must be paid to intra-class variations and external deformations. The Xi'an network inherently supports consistent feature representation and efficient transfer learning for different product SKUs through a dual-path topology and weight sharing.

Batch normalization convolution is the starting point of the structure, which can reduce the variations in lighting and noise caused by different warehouse lighting during deployment. Subsequent layers are organized into multi-scale residual blocks through hierarchical prompts, with each block having a channel attention gate:

$$\mathbf{A}_c = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\mathrm{GAP}(\mathbf{F}_c)\right)\right)\cdot\mathbf{F}_c \quad \text{Eq.(1)}$$

where $\mathbf{F}_c$ is a feature map at channel c, GAP is global average pooling, δ is the ReLU function, and σ is the sigmoid activation. Selectively highlight class-distinguishing features at all depths in this way.

For input images $\mathbf{X}_1$ and $\mathbf{X}_2$, the system produces embeddings $\mathbf{f}_1 = \mathbf{\Phi}(\mathbf{X}_1)$, $\mathbf{f}_2 = \mathbf{\Phi}(\mathbf{X}_2)$. The learned distance metric incorporates global and local distributions:

$$d^* = (\|\mathbf{f}_1 - \mathbf{f}_2\|_2 + \beta\|\mathbf{M}\cdot(\mathbf{f}_1 - \mathbf{f}_2)\|_1)^\alpha \quad \text{Eq.(2)}$$

where $\mathbf{M}$ learns a sparse projection for outlier robustness, and α, β tune the metric's convexity and feature balance.

Observe changes in objectives and profit under supervision:

$$\mathcal{L}_{\text{contrastive}} = (1-y)d^{*2} + y\max(0, m(\mathcal{B}) - d^*)^2 \quad \text{Eq.(3)}$$

where y is the match indicator and $m(\mathcal{B})$ is a per-batch margin updated via reinforcement feedback to penalize ambiguous negatives.

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

The two regularisation terms introduced are for the variance penalty of intra-class dispersion and a general cross-domain orthogonality constraint. To address the problem of dataset imbalance and enhance the discriminatory power of the model:

$$\mathcal{L}_{\text{reg}} = \lambda_1 \cdot \text{Var}[d^* \mid y = 1] + \lambda_2 \cdot \|\langle \mathbf{f}_1, \mathbf{f}_2 \rangle - \mathbf{0}\|_1 \quad \text{Eq.(4)}$$

The following is the end-to-end optimized full network loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{reg}} + \eta \|\text{Cov}(\mathbf{f}_1, \mathbf{f}_2) - \mathbf{I}\|_F^2 \quad \text{Eq.(5)}$$

where η scales the correlation alignment penalty, enforcing consistent geometric embedding distribution.

RMSprop is employed to set a schedule for the backpropagation learning rate. Entropy rate minimisation is used to set the embedding dimension of the high-discrimination retrieval index, and this has been verified. As shown in Figure 1, the final Siamese backbone network will be able to perform stable matching under all operating conditions and uncertainty in real warehouses.

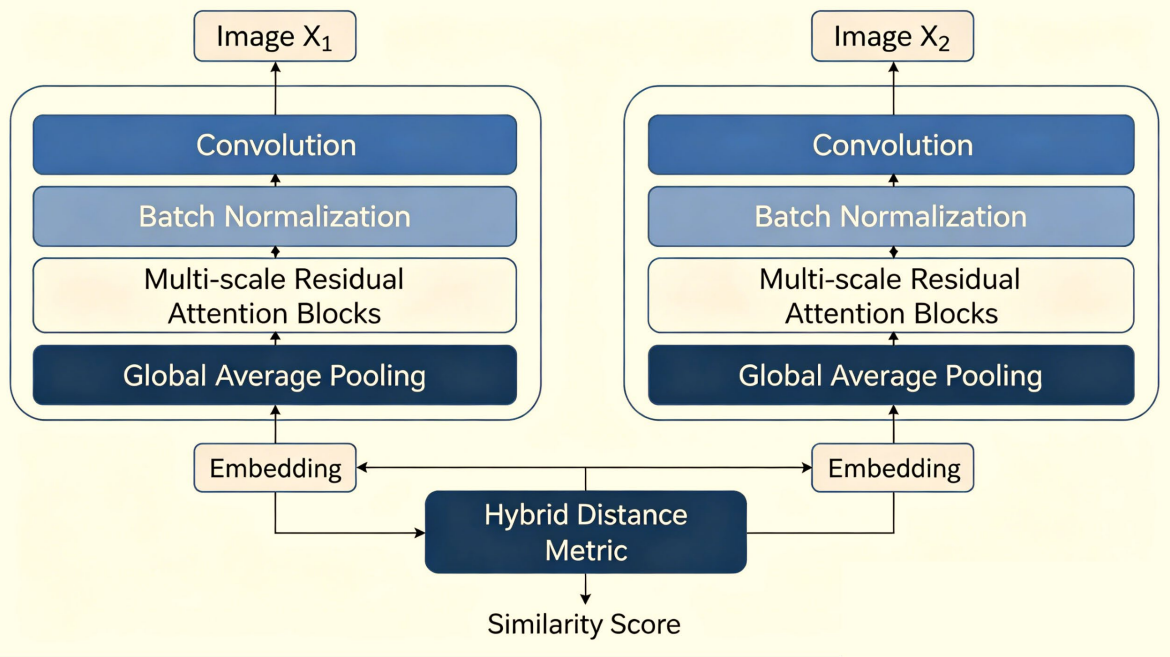

Figure 1. Siamese Network Architecture for Automated Product Visual Matching

Data Augmentation Techniques

For a good generalisation capability of the model, all sorts of data augmentation will be used. Many factors in real-world warehouse environments include dense backgrounds, random occlusions and light variations. The system is not a direct conversion, but a random and context-aware enhancement engine. Create new observation patterns that simulate changes in real work.

Cropping, flipping, and adaptive brightness perturbation are the main parts. To display the transformation sequence, please use image $\mathbf{X}$ input:

$$\mathbf{X}' = \mathcal{T}_{\text{bright}}\left(\mathcal{T}_{\text{flip}}\left(\mathcal{T}_{\text{crop}}(\mathbf{X})\right)\right) \quad \text{Eq.(6)}$$

where $\mathcal{T}_{\text{crop}}$ executes multi-scale region sampling based on real distribution priors, $\mathcal{T}_{\text{flip}}$ randomly mirrors content according to physical placement patterns, and $\mathcal{T}_{\text{bright}}$ imposes nonGaussian color and luminance shifts to simulate onsite illumination disturbances.

Introduced a diversity-induced regularization loss to quantify the improvement in learning stability brought by data augmentation:

$$\mathcal{L}_{\text{div}} = \delta \cdot \sum_{c} \text{Var}\left(\{\Phi(\mathbf{X}_i')\}_{i:y_i=c}\right) \quad \text{Eq.(7)}$$

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

where δ is a tunable factor, Φ is the feature encoder, and the variance is computed over all augmented samples within class c. This module significantly improves the intra-class dispersion in the embedding space. This allows the network to focus on the true invariant product identity, rather than artifacts caused by augmentation.

In practice, a two-level enhancement strategy was used to expand the operational range of the model, and by reducing the risks brought by changes in physical and environmental conditions in large-scale smart warehouses, the accuracy of product matching was improved.

Similarity Computation and Decision Rules

Use Siamese networks to extract representative embeddings, and then create a reliable, context-aware similarity scoring function for identifying products in large-scale warehouses. The system will not use fixed metrics; similarity calculations and matching thresholds will be dynamically adjusted based on class and real-time environmental changes.

For any candidate image pair, let $\mathbf{f}_1$ and $\mathbf{f}_2$ denote their respective embeddings. We introduce an anisotropic hybrid similarity that unifies Mahalanobis and cosine perspectives, capturing both variance-normalized and angular relations:

$$S(\mathbf{f}_1, \mathbf{f}_2) = -\sqrt{(\mathbf{f}_1 - \mathbf{f}_2)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{f}_1 - \mathbf{f}_2)} + \lambda \frac{\mathbf{f}_1^\top \mathbf{f}_2}{\|\mathbf{f}_1\| \|\mathbf{f}_2\|} \quad \text{Eq.(8)}$$

Here, $\boldsymbol{\Sigma}$ is the covariance matrix of embedding vectors adaptively tracked during system operation, and λ balances the influence of directional agreement and distributional similarity. Combining the two can reduce the issue of high-variance pseudo-features while also better preserving the structural information in the embedding space.

Due to the changes in SKU combinations causing the similarity distribution to be very unstable, a self-adjusting threshold has been introduced for real-time matching. For a dynamic validation batch, the optimal threshold τ^* is found by maximizing a regularized Matthews correlation coefficient:

$$\tau^* = \underset{\tau}{\operatorname{argmax}} [\mathrm{MCC}(y, \mathbb{I}(S > \tau)) - \alpha \mathrm{Var}(S)] \quad \text{Eq.(9)}$$

where y is the binary ground truth indicator and α is a regularization constant penalizing excessive threshold sensitivity to batch similarity variance. Can handle high turnover products or continuous changes in other environments.

In order to incorporate the aforementioned parts during the warehouse matching process, the final criteria for selecting the two candidates are:

$$\delta_{\text{match}} = \mathbb{I}(S(\mathbf{f}_1, \mathbf{f}_2) > \tau^*) \cdot \mathbb{I}(\omega(\mathbf{f}_1, \mathbf{f}_2) \geq \eta) \quad \text{Eq.(10)}$$

where $\omega(\mathbf{f}_1, \mathbf{f}_2)$ measures local semantic density around embeddings and η is an empirically calibrated threshold to filter ambiguous pairs, By reducing the congestion of internal space, accuracy has been improved.

As shown in Figure 2, the end-to-end orchestration constitutes the entire workflow of the system, including batch ingestion, visual candidate filtering, Siamese encoding, dynamic similarity measurement, confidence adaptive decision-making, and seamless database updates in the warehouse management platform. Build a feasible and transparent model that can handle large-scale changes occurring between many products.

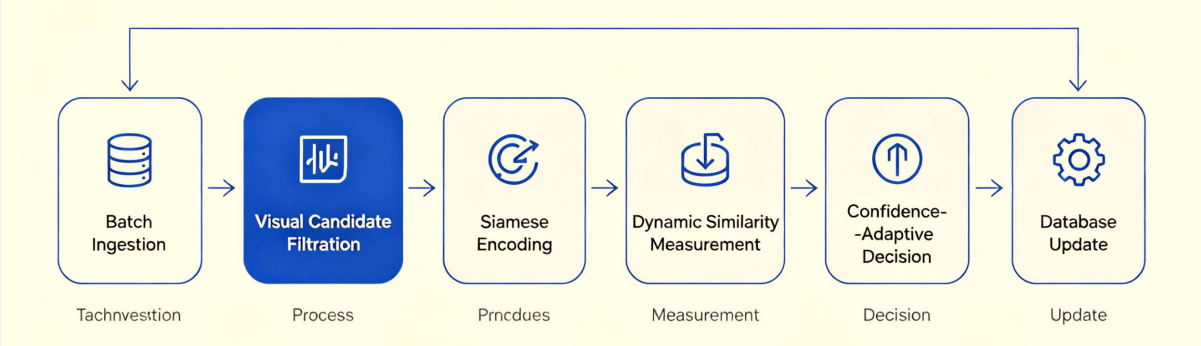

Figure 2. System Workflow and Visual Similarity Computation Pipeline

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

Experimental Evaluation

Dataset Construction and Annotation

In order to verify the reliability of the proposed visual matching framework, a relatively large dataset of various operational conditions in modern warehouses was created. The data was collected at three logistics centers, covering all stages of the receiving, shelving, picking, and shipping processes. High-resolution industrial cameras are placed at different heights and angles to capture product images in both single-stack and multi-stack configurations, under various lighting conditions (including artificial and natural light as well as partial shadows). Product scenes are often chaotic, hidden, and reflective. Labels can sometimes be lost or damaged, and the same issues may occur in actual deployment.

Through expert manual review and calibration of the detector's automatic region suggestions, annotations can be generated. Manually check the bounding box positions and category mappings for each product. In addition to static labels, other attributes have been added, such as acquisition time, storage location, and processing status. The delays in seasonal cycles and other short-term changes will be fine-tuned. To prevent overfitting of the dataset and improve the generalization ability of the evaluation, the dataset was stratified. The training set accounts for 70%, the validation set for 15%, and the test set for 15%. An independent test set will be used, with no SKU overlap in each partition.

Cross-validation uses a strict consensus-based annotation protocol. 15% of all areas underwent double-blind review, and uncommon or ambiguous categories were also confirmed. The annotations meet the high standards of the scalable warehouse visual matching criteria, with very small positional errors.

Evaluation Metrics and Experimental Protocols

It is necessary to collect data on individual work results and overall operational conditions to understand how visual matching performs in large-scale warehouses. Calculate the true positive rate and false positive rate of all evaluation models at the SKU instance level to reflect the actual deployment situation and economic losses due to slight misclassifications.

To evaluate the model's performance in addressing the uneven distribution of warehouse inventory and the open-set characteristics, these metrics are accuracy, precision, recall, and F1-score. Let y_i denote ground truth for the i-th query pair, $\hat{y}_i$ the predicted label, and N the total reference pairs. Overall accuracy is computed as

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \qquad \text{Eq.(11)}$$

where $\mathbb{I}$ is the indicator function. Given the degree of class imbalance, rare SKUs with very few samples coexist with high-frequency products. Micro-average precision and recall are used for a fairer evaluation:

$$\text{Precision} = \frac{\sum_c \text{TP}_c}{\sum_c (\text{TP}_c + \text{FP}_c)}, \text{Recall} = \frac{\sum_c \text{TP}_c}{\sum_c (\text{TP}_c + \text{FN}_c)} \qquad \text{Eq.(12)}$$

where TP_c, FP_c, and FN_c are, respectively, the true positives, false positives, and false negatives for class c.

Due to operational reliability issues caused by severe imbalance, the F1 scores are as follows:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad \text{Eq.(13)}$$

This score is a short number that shows the trade-off between overly conservative (high precision, low recall) and overly lenient (high recall, low precision) predictions. It is crucial to deploy in environments with frequent changes and nearly repetitive packaging.

Set up an experiment to simulate real logistics conditions. Use sequential updates to simulate the addition and deletion of SKUs, and perform cross-validation on all models on non-overlapping folds in both time and space. To test the model's performance under various lighting conditions, levels of clutter, and rare viewpoints, at each time step, the augmentation parameters are based on the observed changes within the facility. Record system throughput, latency, and memory usage to help downstream integration teams quickly assess the practical application of the deployed system. In subsequent optimization rounds, in addition to the aforementioned fuzzy pairs, category outliers and failures in specific scenario clusters were also identified and corrected. Based on a

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

series of evaluations, the increase in metrics is actually the result of improved automated product matching performance.

Ablation Studies

In high-demand warehouse product matching scenarios, ablation studies provide concrete support for the role of system architecture and algorithm improvements and their interrelationships. In order to better understand the functionality of each module, selective disabling or swapping of these modules will be conducted, and the impact on system stability and recognition accuracy will be directly analyzed.

When the residual attention module is not added to the Siamese backbone, its ability to distinguish between various SKUs and its immunity to changes in ambient light decrease. Feature embeddings are more sensitive to background noise and packaging reflections, and the feature maps are often inconsistent with the product features. When channel attention is replaced by static pooling, the problem becomes worse. This makes it more difficult to distinguish products that have the same surface features but differ in functional category or location.

It is expected that during the training process, the types of visual information the model can use will be limited, and the data augmentation pipeline will only use cropping and flipping methods. In this case, the problem is that when the model is deployed in the actual working area, its cross-environment validation accuracy will be very low because these situations did not exist during training. Complete augmentation improves the system's robustness by altering the brightness, shadows, and occlusions in the training distribution, thereby helping the network learn features that are generally invariant.

Fixed, non-adaptive matching thresholds improve the speed of anomaly classification in the similarity module. When new product images are added to the inventory or the SKU turnover rate is high, false positives may occur. Based on the context of local similarity and operational batch characteristics, dynamically adjust the data thresholds to more easily distinguish between matching and non-matching pairs, and adjust decision boundaries, etc. More suitable for large-scale and highly variable environments with frequent changes.

Error analysis indicates that if any major module is excluded, error clusters will concentrate on visually ambiguous or environmentally challenging examples. The entire system is more accurate and responsive in the working environment. Structural reasoning, enhanced reasoning, and adaptive reasoning are all necessary and effective, but a single design innovation cannot solve the warehouse matching problem.

Analysis and Discussion

Impact of Data Variability and Computational Robustness

According to the evaluation of variable warehouse conditions, the accuracy of visual product matching is more sensitive to environmental interference. As shown in Figure 3(a), when the performance curve of the standard model in the gradient-controlled lighting scenario shifts from the sunlit region to the mixed spectrum shadow region, the F1-score decreases by more than 13%. The optimized Siamese framework still maintains over 91% accuracy under severe light gradient conditions, with architecture normalization and targeted enhancement covering the peak irradiance range.

As shown in Figure 3(b), the difficulty of recall under occlusion varies in different areas of the shelf, such as stacking artifacts and partial visibility. In cases of severe occlusion, the error rate of the baseline method is twice that of the low-occlusion shelves. The recall rate and precision only slightly decreased, but due to the impact of the residual attention blocks and the difficult negative sample mining trained on occlusion-rich samples, it remains more robust.

Figure 3(c) shows that performance analysis at different angles is more sensitive to perspective normalization. Traditional models perform poorly at extreme off-center or top-down angles, but the network proposed here maintains embedding separation and category consistency in all tests through geometric jitter enhancement.

As shown in Figure 3(d), the impact of background complexity on the misrecognition rate is relatively small, but not insignificant. Due to the cluttered and noisy storage background, it is difficult to observe. Use customizable masks and context-aware attention to address this issue.

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

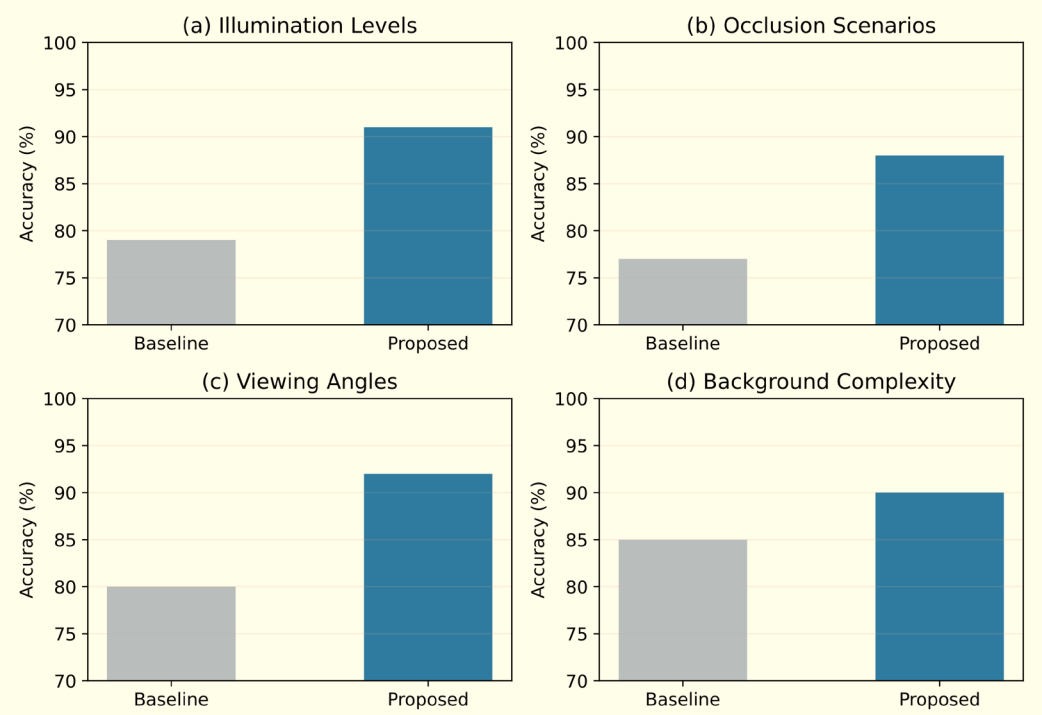

Figure 3. Influence of Data Variability on Matching Accuracy: (a) Illumination levels; (b) Occlusion scenarios; (c) Viewing angles; (d) Background complexity

The computational scalability and accuracy-based analysis of the system have also been validated. Figure 4(a) shows the quantitative inference speeds of embedded ARM devices, edge GPUs, and high-performance warehouse servers. The system achieved a single-pair latency of less than 48 milliseconds on standard hardware, and the total batch throughput improved by over 20% compared to the original Siamese pipeline.

As shown in Figure 4(b), the model size and throughput are related to the model's throughput. The solution is capable of providing real-time responses for models with over 18 million parameters. Until the operational batch size reaches 128 concurrent queries, only in the batch latency curve shown in Figure 4(c) does a nearly linear scaling appear.

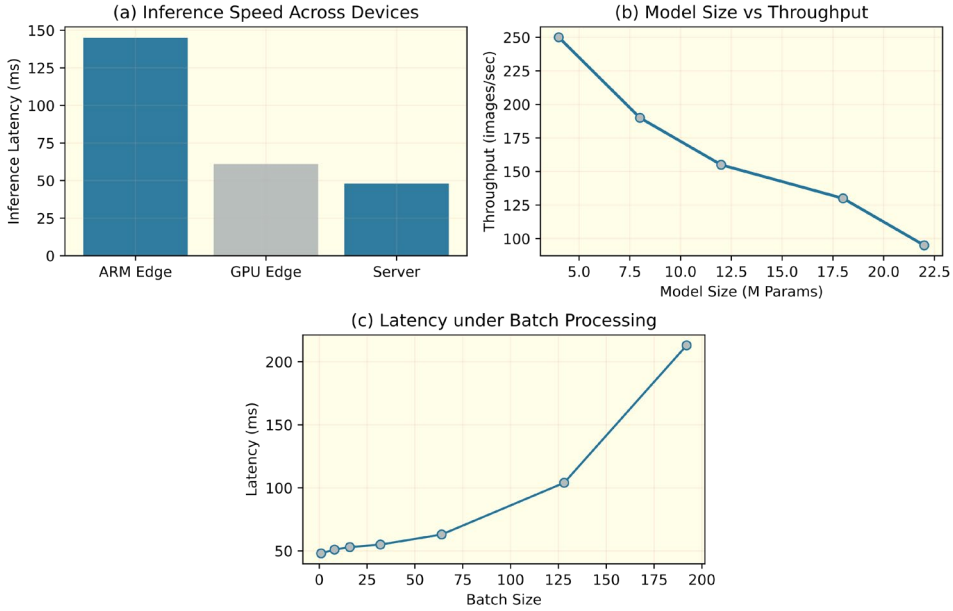

Figure 4. Analysis of Computational Efficiency – Speed and Scalability: (a) Inference speed across devices; (b) Model size vs. throughput; (c) Latency under batch processing

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

The network demonstrates good accuracy and stability in real warehouse environments, meeting the industry's scale and speed requirements while reducing the trade-off between the complexity of traditional algorithms and operational feasibility.

Real-World Performance and Environmental Adaptation

Analyze the on-site deployment of the operational warehouse to improve the applicability and feasibility of the proposed matching system. As shown in Figure 5(a), the category-specific validation for fast-moving consumer goods, durable consumer goods, and other packaging shapes all performed well. The accuracy of most product categories in the model exceeds 90%. Extremely similar visual features or ambiguous SKU cluster labels may reduce accuracy, and even the latest architectures cannot fully resolve this issue.

Figure 5(b) shows how the time robustness model performs at different times of a typical day. Due to context-aware training and time-aware improvements, the Siamese system maintains relatively stable accuracy throughout the day. The baseline drift pattern is the result of decreased output quality due to changes in lighting and working environment.

Figure 5(c) shows the main causes of installation site failures. It also shows the types and relative frequencies of typical error cases. More than two-thirds of the misclassifications are due to the impact of brief occlusions or motion blur on the image sequences during the manual processing; these are not issues of continuous lighting or incorrect labeling. Increase the intervention rate of temporal smoothing and frame-based voting modules in the model to reduce rare but severe errors.

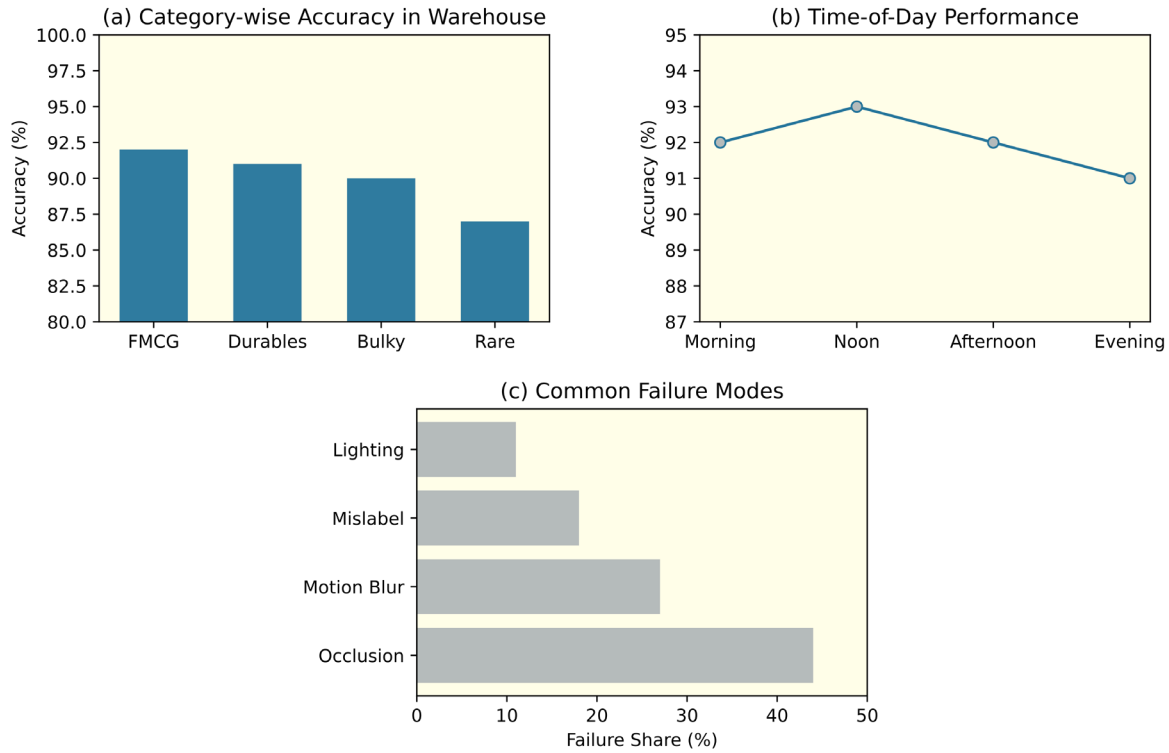

Figure 5. Performance under Real-World Scenarios: (a) Category-wise accuracy in warehouse; (b) Time-of-day performance; (c) Common failure modes

Figure 6(a) shows the environmental transfer experiment inside and outside the warehouse. Both are implemented through a connected network. The various weather-adaptive lighting in the enhanced pipeline improves both recall and accuracy. The normalization module reduces the range of light intensity in open areas.

Figure 6(b) shows the characteristics of extreme temperatures. The competition results will not be affected by summer heat or cold; sensor noise or packaging reflection changes will not cause thermal distortion in the image path.

https://doi.org/10.64972/jaat.2026v4.128pe16:207-219

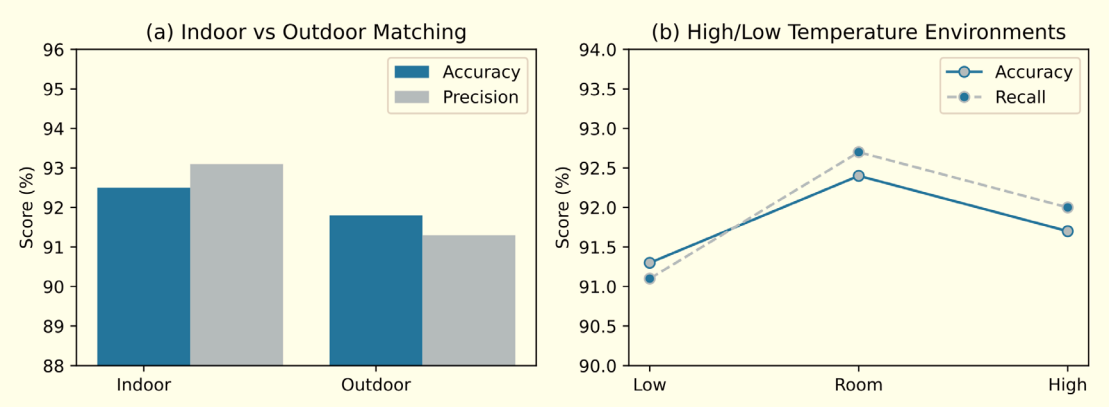

Figure 6. Robustness across Environmental Conditions: (a) Indoor vs outdoor matching; (b) High/low temperature environments

Comparative and Ablation Evaluation

By comparing benchmark cases and ablation experiments, it can be determined that the new data and modules have improved the model's performance. As shown in Figure 7(a), the entire Siamese system outperforms traditional single-stream CNN classifiers and simple distance metric methods. Increase the accuracy by up to 10 percentage points and improve the F1 score in seemingly chaotic and blurry SKU clusters.

As shown in Figures 7(b) and 7(c), removing data augmentation and using a fixed margin objective instead of an adaptive loss function both lead to a decrease in the model's generalization ability and accuracy. If there is no data augmentation, the system will be overly sensitive to changes in lighting and background, and the charts for throughput and error rate will be very unstable. As shown in Figure 7(d), accuracy and efficiency must be matched. Reducing the model's parameters and embedding size will slightly improve inference speed, but at the cost of significantly lowering the actual matching accuracy.

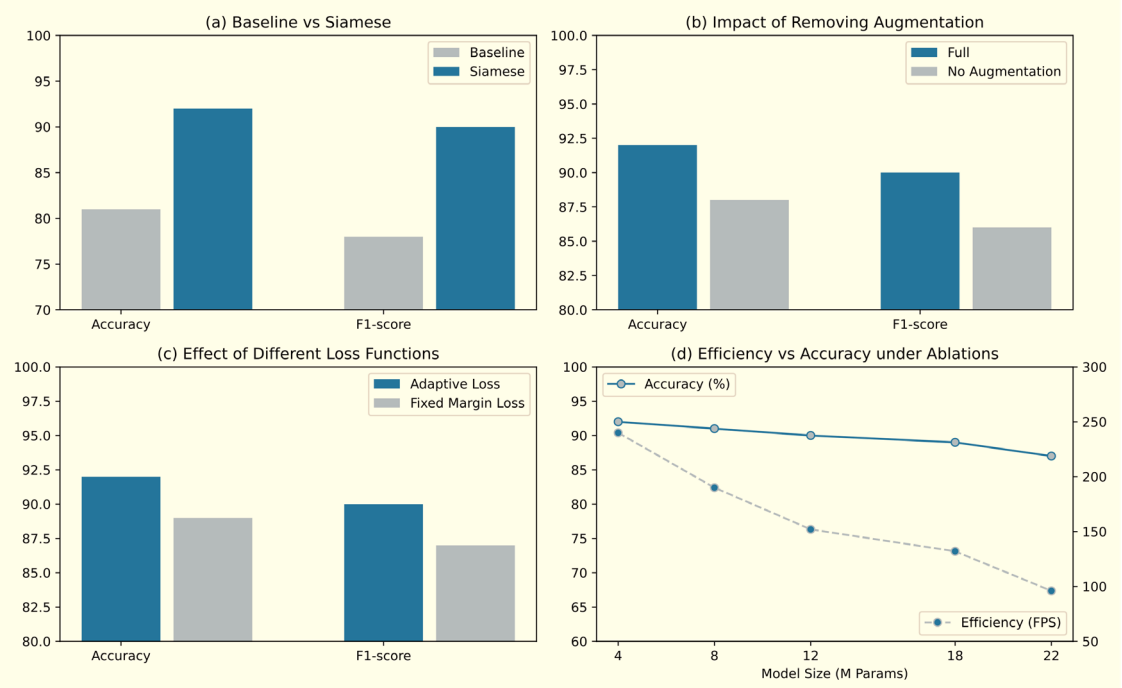

Figure 7. Comparison with Baseline Methods and Ablation Study: (a) Baseline vs Siamese; (b) Impact of removing augmentation; (c) Effect of different loss functions; (d) Efficiency vs accuracy under ablations

The combination of full attention, augmentation, and adaptive optimization arrangements maintains excellent stability while increasing batch size and product diversity. Targeted improvements in representation and training

processes are necessary for robustness in real-life scenarios; without these improvements, performance will decline under adverse deployment conditions.

Conclusion

This paper provides a detailed introduction to a domain-adaptive Siamese network, which is specifically designed for intelligent warehouse environments with complex visual and high operational demands. Compared to other methods, under conditions of partial occlusion, large-scale lighting changes, and cluttered backgrounds, the framework's matching accuracy and robustness are improved by adding residual attention, targeted multi-channel enhancement, and adaptive decision logic. According to empirical tests, it performs better in most cases and is able to handle low visibility and various SKUs well.

The proposed system has good speed and scalability. Real-time inference can still function normally across various devices and distributed endpoints, with different batch sizes and various inventory content. Building a stable, high-performance industrial-grade operating platform will receive support from all modules within the system. The addition of structures, the introduction of new data collection methods, and the adjustment of dynamic thresholds are all included above.

The system is relatively fast and scalable. Real-time inference can still operate normally on various devices and distributed endpoints, and it can provide various types and batches of inventory content. Building a stable, highly reliable industrial-grade operational platform will consist of all the system's modules. These optimizations include architectural improvements, new data collection methods, and adjustable thresholds.

When new issues arise, the framework will continuously expand to provide more automated processing for these new types and distribution changes. The generalization ability of the dataset will be enhanced by adding other factors, such as product shape, packaging materials, and environmental conditions. Single-frame inference can now be extended through temporal modeling and various sensor data streams to handle the dynamic changes occurring throughout the logistics lifecycle. This paper strongly supports the reliability and scalability of visual product matching in next-generation smart warehouse systems.

Author Contributions

Mohamed Al Khalifa contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Aziza Al Romaithi contributes to methodology, software, validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Ma'Rufah, L., Karlita, T., Sa'Adah, U., & Fauzi, W. A. (2023, August). A novel approach to visual search in e-commerce fashion using siamese neural network and multi-scale cnn. In 2023 International Electronics Symposium (IES) (pp. 460-465). IEEE. https://doi.org/10.1109/IES59143.2023.10242507

[2] Yang, Z., Sinnott, R. O., Bailey, J., & Ke, Q. (2023). A survey of automated data augmentation algorithms for deep learning-based image classification tasks. Knowledge and Information Systems, 65(7), 2805-2861. https://doi.org/10.1007/s10115-023-01853-2

[3] Riaz, W., Ullah, A., & Ji, J. (2025). Multi-Scale Attention Networks with Feature Refinement for Medical Item Classification in Intelligent Healthcare Systems. Sensors, 25(17), 5305. https://doi.org/10.3390/s25175305

[4] Olmo, J. J. L. D., Ballesteros, E. P., Gómez, Á. L. P., López-de-Teruel, P. E., Ruiz, A., & Clemente, F. J. G. (2025). Fog computing-driven logistics: leveraging few-shot learning and foundational computer vision models. Cluster Computing, 28(14), 938. https://doi.org/10.1007/s10586-025-05662-w

[5] Zhao, J., Zhan, W., Zhao, W. X., Zhang, Q., Gui, T., Wei, Z., ... & Sun, M. (2023, July). Re-matching: A fine-grained semantic matching method for zero-shot relation extraction. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6680-6691). https://doi.org/10.18653/v1/2023.acl-long.369

[6] Godwill, J. (2025). Explainable Multimodal Product Classification: Interpreting YOLO-OCR Fusion Decisions for Transparent Retail AI Systems. Available at SSRN 5575710. http://dx.doi.org/10.2139/ssrn.5575710

[7] Yang, J., & Lee, C. H. (2025, April). Real-Time Data-Driven Method for Bolt Defect Detection and Size Measurement in Industrial Production. In Actuators (Vol. 14, No. 4, p. 185). MDPI. https://doi.org/10.3390/act14040185

[8] Li, S., Lv, F., Jin, T., Lin, G., Yang, K., Zeng, X., ... & Ma, Q. (2021, August). Embedding-based product retrieval in taobao search. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 3181-3189). https://doi.org/10.1145/3447548.3467101

[9] Vu, V. D., Hoang, D. D., Tan, P. X., Nguyen, V. T., Nguyen, T. U., Hoang, N. A., ... & Hoang, D. C. (2024). Occlusion-robust pallet pose estimation for warehouse automation. IEEE Access, 12, 1927-1942. https://doi.org/10.1109/ACCESS.2023.3348781

[10] Manoj, K. C., & Dhas, D. A. S. (2022). Automated brain tumor malignancy detection via 3D MRI using adaptive-3-D U-Net and heuristic-based deep neural network. Multimedia Systems, 28(6), 2247-2273. https://doi.org/10.1007/s00530-022-00952-4

[11] Liu, H., Zhou, L., Zhao, J., Wang, F., Yang, J., Liang, K., & Li, Z. (2022). Deep-learning-based accurate identification of warehouse goods for robot picking operations. Sustainability, 14(13), 7781. https://doi.org/10.3390/su14137781

[12] Fan, C., Yu, H., Huang, Y., Shan, C., Wang, L., & Li, C. (2021). SiamON: Siamese occlusion-aware network for visual tracking. IEEE Transactions on Circuits and Systems for Video Technology, 33(1), 186-199. https://doi.org/10.1109/TCSVT.2021.3102886

[13] Li, S., Guo, H., Tang, X., Tang, R., Hou, L., Li, R., & Zhang, R. (2024). Embedding compression in recommender systems: A survey. ACM Computing Surveys, 56(5), 1-21. https://doi.org/10.1145/3637841

[14] Wen, X., Zhao, B., Zheng, A., Zhang, X., & Qi, X. (2022). Self-supervised visual representation learning with semantic grouping. Advances in neural information processing systems, 35, 16423-16438. https://doi.org/10.58496/BJML/2025/004

[15] Huang, C., Jiang, S., Li, Y., Zhang, Z., Traish, J., Deng, C., ... & Da Xu, R. Y. (2020, August). End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. In European Conference on Computer Vision (pp. 477-493). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58604-1_29

[16] Liu, A. A., Guo, F. B., Zhou, H. Y., Yan, C. G., Gao, Z., Li, X. Y., & Li, W. H. (2022). Domain-adversarial-guided Siamese network for unsupervised cross-domain 3-D object retrieval. IEEE Transactions on Cybernetics, 52(12), 13862-13873. https://doi.org/10.1109/TCYB.2021.3139927

[17] Shu, X., Zhang, L., Wang, Z., Wang, L., & Yi, Z. (2023). Fine-grained recognition: Multi-granularity labels and category similarity matrix. Knowledge-Based Systems, 273, 110599. https://doi.org/10.1016/j.knosys.2023.110599

[18] Hu, Q., Guo, Y., Cordy, M., Xie, X., Ma, L., Papadakis, M., & Le Traon, Y. (2022). An empirical study on data distribution-aware test selection for deep learning enhancement. ACM Transactions on Software Engineering and Methodology (TOSEM), 31(4), 1-30. https://doi.org/10.1145/3511598

[19] Yin, H., Chen, C., Hao, C., & Huang, B. (2022). A Vision-based inventory method for stacked goods in stereoscopic warehouse. Neural computing and applications, 34(23), 20773-20790. https://doi.org/10.1007/s00521-022-07551-4

[20] Di Capua, M., Ciaramella, A., & De Prisco, A. (2023). Machine learning and computer vision for the automation of processes in advanced logistics: The integrated logistic platform (ILP) 4.0. Procedia Computer Science, 217, 326-338. https://doi.org/10.1016/j.procs.2022.12.228

[21] van Geest, M., Tekinerdogan, B., & Catal, C. (2021). Smart warehouses: Rationale, challenges and solution directions. Applied sciences, 12(1), 219. https://doi.org/10.3390/app12010219

[22] Jamshed, A., Mallick, B., & Kumar, P. (2020). Deep learning-based sequential pattern mining for progressive database: A. Jamshed et al. Soft Computing, 24(22), 17233-17246. https://doi.org/10.1007/s00500-020-05015-2

[23] Shi, H., Liu, M., Mu, X., Song, X., Hu, Y., & Nie, L. (2024). Breaking through the noisy correspondence: A robust model for image-text matching. ACM Transactions on Information Systems, 42(6), 1-26. https://doi.org/10.1145/3662732

[24] Qu, S., & Hu, G. (2024). Scalable learning for multiagent route planning: Adapting to diverse task scales. IEEE Transactions on Artificial Intelligence, 5(10), 4996-5011. https://doi.org/10.1109/TAI.2024.3402193

[25] Weyns, D., Bäck, T., Vidal, R., Yao, X., & Belbachir, A. N. (2023). The vision of self-evolving computing systems. Journal of Integrated Design and Process Science, 26(3-4), 351-367. https://doi.org/10.3233/JID-220003