IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5736~5744
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5736-5744

Classification of industrial products using vision transformer architecture

Robinson Jiménez-Moreno[1], Anny Astrid Espitia-Cubillos[2], Esperanza Rodríguez-Carmona[2]
[1]Department of Mechatronic Engineering, Faculty of Engineering, Universidad Militar Nueva Granada, Bogotá, Colombia
[2]Department of Industrial Engineering Program, Faculty of Engineering, Universidad Militar Nueva Granada, Bogotá, Colombia

Article Info

Article history:

Received Dec 2, 2025
Revised Aug 14, 2026
Accepted Sep 20, 2026

Keywords:

Industry application
Labeling
Transformer networks
Vision transformer architecture
You only look once architecture

ABSTRACT

Recognition of patterns in images by means of neural networks (NN) is becoming more specialized, where initially preprocessing was required now, by means of deep learning algorithms, it is possible to obtain identification and location of objects and their characteristics. This paper presents the use of transformer vision NN and their comparison with you only look once (YOLO) networks, in the framework of Industry 5.0, to recognize and label industrial products with similar packaging, which increases complexity of differentiation. Process begins with construction of a database of 25 images of each product with different backgrounds. Two types of networks are trained, defining learning parameters. Validated networks are tested under normal conditions, with images with low level of detail and backlighting, with images that are not centered, incomplete, of untrained products, and inside shelves, which allowed for identification and correction of faults. Effectiveness was validated, obtaining for vision transformer (ViT) network an accuracy of 100% with an average classification time of 0.4302 seconds and for YOLO V4 network an accuracy of 75% with a classification time of 0.0254 seconds. The use of the ViT network is recommended for its high accuracy due to its overall robustness.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Anny Astrid Espitia-Cubillos
Department of Industrial Engineering Program, Faculty of Engineering, Universidad Militar Nueva Granada
Carrera 11 #101-80, Bogotá, Colombia
Email: anny.espitia@unimilitar.edu.co

1. INTRODUCTION

The inclusion of artificial intelligence (AI) in the production chain is oriented. According to product standardization, one of the activities susceptible to automation is product identification, which allows for real-time measurement of productivity, inventory levels of each product, its location, among others, facilitating timely, informed decision-making in production and logistics processes. Managing inventory and measuring productivity in real time is difficult when products are visually similar.

Neural networks (NN) that are currently at the forefront of knowledge, such as transformer networks [1] are revolutionizing image-based pattern recognition systems. Nowadays recognized as vision transformer (ViT), these network architectures present high performance in feature extraction [2] in image-based classification [3]. Where its applications range from detection of falsifications in audio and video [4], classification in electroencephalogram (EEG) images [5], vision defect detection in manufacturing processes [6], automatic cancer detection systems [7], histopathological image analysis [8], 3D tumor segmentation [9], among many others.

Thus, its use has been integrated into a variety of systems such as reasoning modules [10], precision agriculture systems [11], and high-resolution feature extraction in satellite images [12]. Being part of hybrid

algorithms with convolutional networks also based on deep learning [13]–[15] and its derivatives such as the you only look once (YOLO) network [16]–[19]. Where a high level of accuracy is achieved even under a small number of data [20], pre-trained networks [21] and through transfer learning techniques [22]. convolutional neural networks (CNNs) exhibit inductive biases that ViTs do not [11]–[13]; however, ViT has quadratic computational complexity that hinders the processing of high-resolution images [11]; in response, a lightweight architecture like the shifted window (Swin) transformer introduces a hierarchical structure that reduces complexity to a linear level [4], [13], [20], data-efficient image transformer (DeiT) employs knowledge distillation to improve efficiency on small-scale industrial datasets [2], and the hybrid window sequence merging-based vision transformer (WSM-ViT) model improves learning without massive pre-training [3]. The limited capability of conventional image processing techniques and CNNs to recognize products with similar appearances is identified as a research gap; attempts to address this have involved complex hybrid models, such as CNN + speeded-up robust features (SURF), in data-scarce scenarios requiring robust models aligned with the needs of modern industrial logistics.

Given the trends of Industry 4.0 [23] and 5.0 [24], where algorithms such as those based on ViT project a great applicability according to the exposed state of the art. In the following, the use of robust architectures such as ViT [25], [26] or YOLO [18], [19] is presented by means of transfer of learning, both applied to the identification of products with application to an industrial cleaning company, where the complexity in the discrimination given the variety of products with the same type of packaging, but identified only by personalization on the label generates a high level of difficulty for image processing techniques [27] and convolutional networks [28], previously worked on.

2. METHOD

Six phases were defined for the development of this study. The first step corresponds to the selection of the products to be identified; in the second instance is to organize the digital images of the products in a database, thirdly, the parameters are defined, and two network architectures are trained, after which each model is validated in different scenarios and the errors identified are corrected by adjusting database that is validated until the results obtained are satisfactory. Finally, the results of the two models considered are compared.

To identify the 15 categories of products with the highest demand in a cleaning supplies manufacturing company, where 11 of them correspond to products with the same type of packaging. A database is built where images of the front view of the label and side views are included to allow from 50% of its visibility and the background of the product is changed so that learning is focused on the characteristics of each element independent of the background. Similarly, the resolution of the images varies in relation to the focal length of the camera; in some cases, the text on the label is not clearly legible, making recognition difficult. This is why conventional architectures have not been successful, requiring for example the integration of techniques, such as CNN and SURF [25]. Figure 1 illustrates an example of the 15 categories chosen, focusing on the classes most likely to present confusion, showing pairs of 8 products in the same package with background changes between each pair. For the present research, 25 images per class are used to validate the robustness of the networks to be tested. The other four categories correspond to products of the same class but packed in containers of greater or lesser capacity, as shown in Figure 2. Finally, 72% training images, 16% validation images and 12% test images are used.

Figure 3 illustrates the flow diagram training transformer networks for the identification of products that have the same type of packaging but differ in labeling and that belong to the inventory of an industrial cleaning company. To prevent overfitting, pre-trained networks are used, training hyperparameters are kept fixed, and the dataset is iteratively improved based on validation and real-world tuning results; this leads to the inclusion of data featuring varying background colors, images with centered and off-center products, occluded labels, different cap colors, and diverse lighting conditions. The entire process was performed on an Intel Core i9 laptop with 24 GB of RAM and an NVIDIA GeForce RTX 4080 graphics card.



Figure 1. Sample database



Figure 2. Sample by-products database

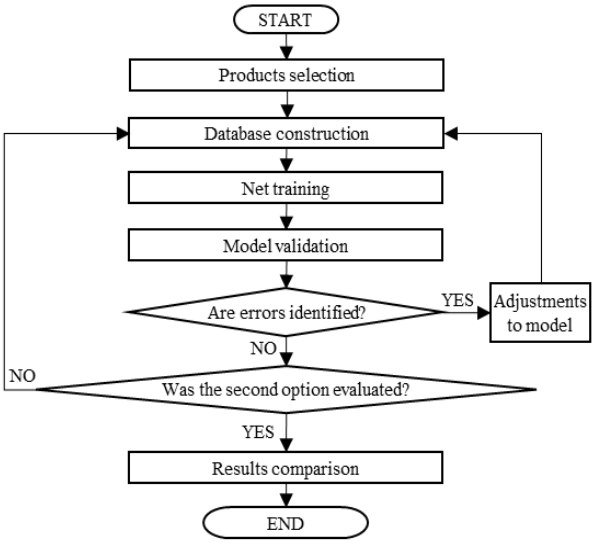

Figure 3. Procedure flowchart

3. RESULTS AND DISCUSSION

3.1. Vision transformer network

The ViT network used has 143 layers and 86.8 million learning parameters and has an input image of 384 × 384 × 3 pixels. Table 1 illustrates the parameters defined for training, which were derived from the database requires a small mini-batch size. The input volume of the network required a previous resizing of the images to a size of 348 pixels per side, as color red, green, blue (RGB) images. Given the predefined architecture of the network, which has 143 layers of depth and is quite robust, few training epochs are used, with a learning rate of 0.0001. The training took 19 minutes and 59 seconds, took 2,340 iterations, and presented the behavior reflected in Figure 4.

Table 1. ViT network training parameters

Parameter	Input volume	Mini-batch size	Optimizer	Epochs	Learning rate
Value	384 × 384 × 3	2	ADAM	20	0.0001

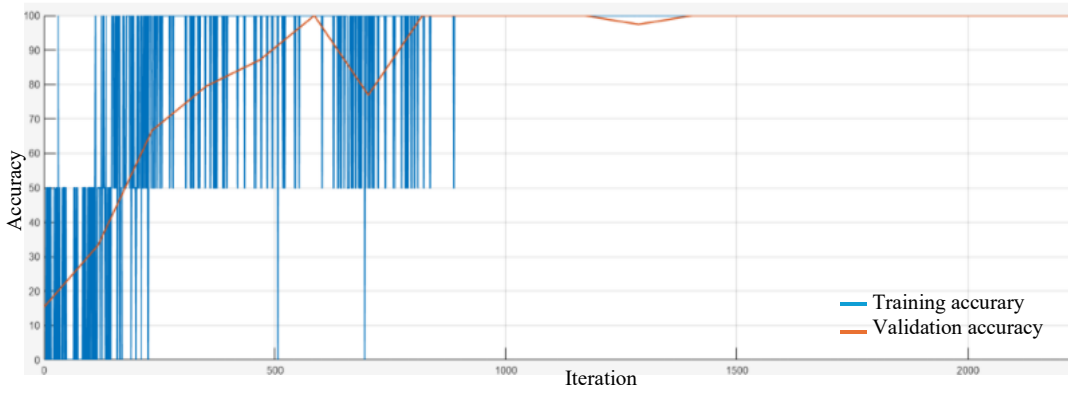

Figure 4. Progress in ViT network training

Figure 5 shows the confusion matrix obtained after training, which presents an accuracy of 100%, without any confusion between the set of test images (4 per category) in all the classes considered. The low number of images used highlights the high capacity of the network in the identification of each class with the learning conditions required by the database with recall, sensitivity, and F1-score results of 100%. Subsequently, the network is tested, and the successful validation results are presented in Figure 6. This allows us to calculate the average classification time, which corresponds to 0.4302 seconds. This highlights that, even with a very deep architecture such as the one used, the response time is sufficient for a product handling job by automatic means on production lines, whether for dispatching, inventory or storage. Figure 7 illustrates how for training samples with a low level of detail and backlighting, where the camera points upwards since the product is on a high shelf, a correct validation result is obtained.

Figure 8 illustrates how even though the camera image is not centered and part of the label is occluded, the class is correctly identified, the blue product that does not allow reading of the label is an untrained product that is detected in the cleaner class, on the right are products of this class with small packaging in two color presentations that are correctly classified.

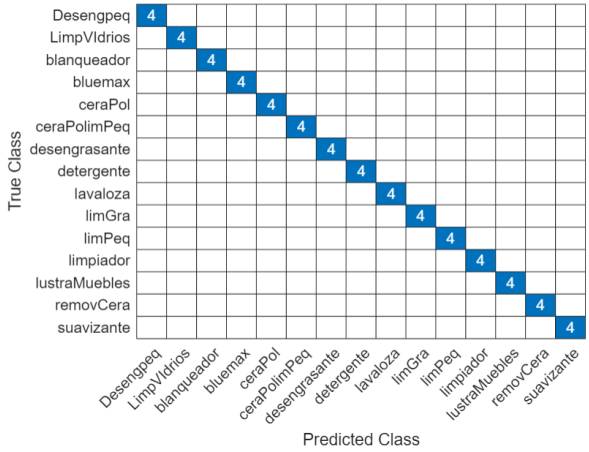

Figure 5. Confusion matrix



Figure 6. ViT validation results



Figure 7. Backlight validation results



Figure 8. Cleaner class results

Figure 9 illustrates the detection of the furniture polish class erroneously when trying to detect it on the shelf, at two different chamber depths, in both cases generating classification error due to blurring. When validating the cause of the error it is found in the database the class lustra furniture that no sample has white cap as the test image on site, as shown in Figure 10. Therefore, database and results adjustment of this class were required, this is show in Figure 11, were database is adjusted (Figure 11(a)) and the network is retrained, solving the classification error as illustrated in Figure 11(b).



Figure 9. Furniture polish class results



Figure 10. Furniture polish class results



(a) (b)

Figure 11. Database and results adjustment for the furniture polish class (a) adjusted database (b) corrected classification results

3.2. You only look once network

To compare the results of the ViT network, a robust architecture based on CNN networks, such as YOLO is used; for this case, version 4 is used, to make the results comparable, the same final database used for the training of the ViT network is used, but the labeling is modified since it must be by location boxes, as illustrated in Figure 12. Table 2 lists the training parameters used for training the YOLO network. The input volume required resizing the images to a size of 416 pixels per side, implying a depth of 3 for each color component. A mini-batch-size of 4 is selected. 80 training epochs are used, with a learning rate of 0.001. The training took 14 minutes and 51 seconds.

The transfer learning strategy relied on the Tiny-YOLOv4 architecture pre-trained on the common objects in context (COCO) dataset, available in MATLAB's computer vision toolbox. Instead of randomly

initializing the network weights, the convolutional backbone, already trained on COCO and capable of extracting generic low and mid-level features like edges, textures and shapes were reused. The final detection layers were replaced with heads tailored to the number of classes in the cleaning products domain (13 classes) and to the dataset-specific anchors. This approach significantly reduces the number of epochs and the amount of annotated data required to achieve adequate convergence. Table 3 shows the transfer learning parameters. The YOLO network is then tested with the same test set as the ViT (Figure 6).



Figure 12. Sample YOLO database labeling

Table 2. YOLO network training parameters

Parameter	Input volume	Mini-batch size	Optimizer	Epochs	Learning rate	Number of classes
Value	416 × 416 × 3	4	ADAM	80	0.001	15

Table 3. YOLO transfer learning parameters

Backbone	Neck	Head	Weights	Size
CSPDarknet53	Feature pyramid network	Two scales (13× 13 and 26× 26)	COCO pre-trained weights: fine-tuned using the common objects in context dataset	23 MB

The successful validation results correspond to 75% with recall, sensitivity, mAp and F1-score results of 92.3%, 92.3%, 92.3% and 72%, respectively, the value per class is presented in Figure 13. In Figure 14 the classified products are presented with the label and the location box, the others were not recognized by the network. The average classification time was 0.025383 seconds. This highlights the improved performance of the ViT network in terms of accuracy, despite having a longer processing time. Continuing the comparison with the same images, Figure 15 illustrates how, unlike the ViT network (Figure 8), the YOLO network does not recognize most of the images in the real scenario. In the case of the comparison with Figure 7 with a backlit image (Figure 16), the classification was successful, but with shadows, as in the case of the furniture polish in Figure 9, it fails to recognize it, even having used the adjusted database for training the YOLO network.

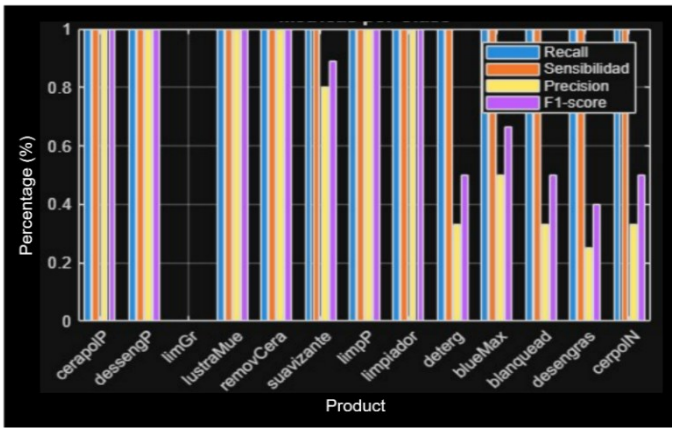

Figure 13. Metrics by class

The results obtained are in contrast to previous works [27] that allow product classification with AI requiring hybrid models for database recognition such as the one presented, which highlights the high impact of the use of transformer networks in applications oriented to the industry, with sufficient time for the manipulation of the identified product. In this sense, [29] point out that, in order to select the most convenient

model according to the end user's objectives in an industrial production environment [30], training times and, in this case, the depth of the network are not relevant, but criteria such as accuracy and classification times are relevant. A limitation of the present development is the size of the dataset (25 images per class). The work is distinguished by its focus on robustness against extreme visual similarity and its efficiency in Industry 5.0 scenarios.



Figure 14. YOLO validation results



Figure 15. Cleaning class results



Figure 16. Backlight comparison

4. CONCLUSION

The results demonstrate that, for robust transformer architectures, an extensive database is not required, even when the input data presents classification challenges like those used in this study, thanks to their pre-training. Despite the depth of the ViT network, with the defined training criteria, the network training took less than 20 minutes, with processing times of less than 0.45 seconds, which is useful for automatic applications that require fast implementation and execution. What is even more interesting is that a 100% accuracy level was reached, after only a few epochs. On the other hand, the YOLO network took less time to train (less than 15 minutes) and has a processing time of less than 0.03 seconds, which makes it very fast; however, its accuracy is equivalent to 75% of the cases tested. Given these results, the use of the ViT network is recommended due to its high precision, considering applications in order picking that require adequate logistic management to avoid claims or returns from customers and reprocesses in order management involving reverse logistics aligned with Industry 5.0. Tests were carried out on both networks in different scenarios and varied conditions, which allowed not only the identification of faults but also their correction to have robust networks. This exhaustive validation of the trained networks and their improvement constitute an iterative process that allows their optimization that can be replicated in various scenarios. Not only was the automation of product identification achieved to facilitate real-time traceability, but a robust tool is also provided that reduces errors, thereby contributing to customer satisfaction. Given the computational complexity of ViT and the longer classification times involved in deploying it to real-time production environments with limited resources, future research will evaluate variants such as the Swin transformer or hybrid models utilizing edge computing, including k-fold cross-validation to reduce the risk of overfitting and ablation studies.

ACKNOWLEDGMENTS

The authors would like to thank the Universidad Militar Nueva Granada, from where they are full-time associate professors.

FUNDING INFORMATION

Product derived from the research project titled “Fortalecimiento de procesos de recepción de pedidos y control de inventario de materias primas soportado en Industria 4.0” INV-ING-4150 financed by the vice-rector for research of the Universidad Militar Nueva Granada, year 2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Robinson Jiménez-Moreno	✓	✓		✓	✓	✓			✓					
Anny Astrid Espitia-Cubillos	✓			✓						✓	✓	✓	✓	✓
Esperanza Rodríguez-Carmona			✓		✓		✓	✓	✓					

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [AAEC], upon reasonable request.

REFERENCES

[1] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, 2017, doi: 10.1201/9781003561460-19.

[2] O. Elharrouss *et al.*, "ViTs as backbones: leveraging vision transformers for feature extraction," *Information Fusion*, vol. 118, 2025, doi: 10.1016/j.inffus.2025.102951.

[3] E. Jiao, Q. Leng, J. Guo, X. Meng, and C. Wang, "Vision transformer with window sequence merging mechanism for image classification," *Applied Soft Computing*, vol. 171, 2025, doi: 10.1016/j.asoc.2025.112811.

[4] S. Usmani, S. Kumar, and D. Sadhya, "Spatio-temporal knowledge distilled video vision transformer (STKD-VViT) for multimodal deepfake detection," *Neurocomputing*, vol. 620, 2025, doi: 10.1016/j.neucom.2024.129256.

[5] Balendra, P. C. Negi, N. Sharma, and S. Sharma, "Scalogram sets based motor imagery EEG classification using modified vision transformer: a comparative study on scalogram sets," *Biomedical Signal Processing and Control*, vol. 104, 2025, doi: 10.1016/j.bspc.2025.107640.

[6] X. Zhang, X. Liang, Y. Zhang, J. Li, and S. Wei, "MLR-WM-ViT: global high-performance classification of mixed-type wafer map defect using a multi-level relay vision transformer," *Expert Systems with Applications*, vol. 277, 2025, doi: 10.1016/j.eswa.2025.127121.

[7] E. C.-Martos, J. V.-Francés, J. V. Bagan, and Y. V.-Gilabert, "Automated classification of oral cancer lesions: vision transformers vs radiomics," *Computers in Biology and Medicine*, vol. 189, 2025, doi: 10.1016/j.compbiomed.2025.109913.

[8] A. N. Mir, D. R. Rizvi, and M. R. Ahmad, "Enhancing histopathological image analysis: an explainable vision transformer approach with comprehensive interpretation methods and evaluation of explanation quality," *Engineering Applications of Artificial Intelligence*, vol. 149, 2025, doi: 10.1016/j.engappai.2025.110519.

[9] D. Pan, J. Shen, Z. Al-Huda, and M. A. A. Al-qaness, "VcaNet: vision transformer with fusion channel and spatial attention module for 3D brain tumor segmentation," *Computers in Biology and Medicine*, vol. 186, 2025, doi: 10.1016/j.compbiomed.2025.109662.

[10] C. Dewi, H. P. Chernovita, S. A. Philemon, C. A. Ananta, and A. P. S. Chen, "Adjusted reasoning module for deep visual question answering using vision transformer," *Computers, Materials and Continua*, vol. 81, no. 3, pp. 4195–4216, 2024, doi: 10.32604/cmc.2024.057453.

[11] S. Ulukaya and S. Deari, "A robust vision transformer-based approach for classification of labeled rices in the wild," *Computers and Electronics in Agriculture*, vol. 231, 2025, doi: 10.1016/j.compag.2025.109950.

[12] J. Bolcek *et al.*, "A comprehensive evaluation of deep vision transformers for road extraction from very-high-resolution satellite data," *Science of Remote Sensing*, vol. 11, 2025, doi: 10.1016/j.srs.2024.100190.

[13] X. Zou, Y. Peng, G. Li, and X. Cao, "PCSViT: efficient and hardware friendly pyramid vision transformer with channel and spatial self-attentions," *Neurocomputing*, vol. 636, 2025, doi: 10.1016/j.neucom.2025.129987.

[14] S. Vaddadi, V. Srinivas, N. A. Reddy, H. Girish, D. Rajkiran, and A. Devipriya, "Factory inventory automation using Industry 4.0 technologies," in *2022 IEEE IAS Global Conference on Emerging Technologies, GlobConET 2022*, 2022, pp. 734–738, doi: 10.1109/GlobConET53749.2022.9872416.

[15] B. Wu, "Motion control algorithm for automatic welding of complex intersecting line joints based on deep learning," in *Proceedings - 2023 International Conference on Mechatronics, IoT and Industrial Informatics, ICMIII 2023*, 2023, pp. 352–356, doi: 10.1109/ICMIII58949.2023.00073.

[16] X. Yang, E. D. R. Castillo, Y. Zou, L. Wotherspoon, J. Yang, and H. Li, "Automated concrete bridge damage detection using an efficient vision transformer-enhanced anchor-free YOLO," *Engineering*, vol. 51, pp. 311–326, 2025, doi: 10.1016/j.eng.2025.02.018.

[17] A. M. Qureshi *et al.*, "Semantic segmentation and YOLO detector over aerial vehicle images," *Computers, Materials and Continua*, vol. 80, no. 2, pp. 3315–3332, 2024, doi: 10.32604/cmc.2024.052582.
[18] Y. Qi and H. Sun, "Defect detection of insulator based on YOLO network," in *2024 9th International Conference on Electronic Technology and Information Science, ICETIS 2024*, 2024, pp. 232–235, doi: 10.1109/ICETIS61828.2024.10593675.
[19] H. Zheng, X. Chen, H. Cheng, Y. Du, and Z. Jiang, "MD-YOLO: surface defect detector for industrial complex environments," *Optics and Lasers in Engineering*, vol. 178, 2024, doi: 10.1016/j.optlaseng.2024.108170.
[20] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-wise convolutions in vision transformers for efficient training on small datasets," *Neurocomputing*, vol. 617, 025, doi: 10.1016/j.neucom.2024.128998.
[21] L. Huang, J. Zeng, M. Yu, W. Ding, X. Bai, and K. Wang, "Efficient feature selection for pre-trained vision transformers," *Computer Vision and Image Understanding*, vol. 254, 2025, doi: 10.1016/j.cviu.2025.104326.
[22] B. Zhang, L. Wang, C. Zhang, R. Zhao, and J. Sun, "No-reference image quality assessment based on improved vision transformer and transfer learning," *Signal Processing: Image Communication*, vol. 135, 2025, doi: 10.1016/j.image.2025.117282.
[23] F. E. Habib, H. Bnouachir, M. Chergui, and A. Ammoumou, "Industry 4.0 concepts and implementation challenges: literature review," in *Proceedings - 2022 9th International Conference on Wireless Networks and Mobile Communications, WINCOM 2022*, 2022, doi: 10.1109/WINCOM55661.2022.9966456.
[24] L. Monferdini, L. Tebaldi, and E. Bottani, "From Industry 4.0 to Industry 5.0: opportunities, challenges, and future perspectives in logistics," *Procedia Computer Science*, vol. 253, pp. 2941–2950, 2025, doi: 10.1016/j.procs.2025.02.018.
[25] A. Dosovitskiy *et al.* "An image is worth 16 x 16 words: transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021.
[26] H. Touvron, M. Cord, A. El-Nouby, J. Verbeek, and H. Jégou, "Three things everyone should know about vision transformers," in *Computer Vision–ECCV 2022*, Cham, Switzerland: Springer, 2022, pp. 497–515, doi: 10.1007/978-3-031-20053-3_29.
[27] S. Chen, D. Liu, Y. Pu, and Y. Zhong, "Advances in deep learning-based image recognition of product packaging," *Image and Vision Computing*, vol. 128, 2022, doi: 10.1016/j.imavis.2022.104571.
[28] L. Kumar, R. Dutt, and K. K. Gaikwad, "Packaging 4.0: artificial intelligence and machine learning applications in the food packaging industry," *Current Food Science and Technology Reports*, vol. 3, no. 19, 2025, doi: 10.1007/s43555-025-00064-w.
[29] A. A. E. Cubillos, R. J.-Moreno, and E. R. Carmona, "Deep learning architectures for location and identification in storage systems," *IAES International Journal of Artificial Intelligence*, vol. 14, no. 1, pp. 592–601, 2025, doi: 10.11591/ijai.v14.i1.pp592-601.
[30] W. Ruangsang, T. N. Nan, and S. Aramvith, "A hybrid human-in-the-loop framework for industrial cocoa bean grading using YOLOv11 and vision transformers," *IEEE Access*, vol. 14, pp. 63480–63493, 2026, doi: 10.1109/ACCESS.2026.3685873.

BIOGRAPHIES OF AUTHORS



Robinson Jiménez-Moreno is an electronic engineer graduated from Universidad Distrital Francisco José de Caldas in 2002. He received a M.Sc. in Engineering from Universidad Nacional de Colombia in 2012 and Ph.D. in Engineering at Universidad Distrital Francisco José de Caldas in 2018. His current working as associate professor of Universidad Militar Nueva Granada and research focuses on the use of CNNs for object recognition and image processing for robotic applications, such as human-machine interaction. He can be contacted at email: robinson.jimenez@unimilitar.edu.co.



Anny Astrid Espitia-Cubillos performed her undergraduate studies in Industrial Engineering in the Universidad Militar Nueva Granada in 2002 and M.Sc. in Industrial Engineering from the Universidad de Los Andes in 2006. She is an associate professor on Department of Industrial Engineering Program at Universidad Militar Nueva Granada, Bogotá, Colombia. She can be contacted at email: anny.espitia@unimilitar.edu.co.



Esperanza Rodríguez-Carmona performed her undergraduate studies in Mechanical Engineering (1997) from the Universidad Tecnológica de Pereira and Master's degree in university teaching from the Universidad de La Salle, Bogotá, Colombia. She is an associate professor on Department of Industrial Engineering program at Universidad Militar Nueva Granada, Bogotá, Colombia. She can be contacted at email: esperanza.rodriguez@unimilitar.edu.co.