A Contrastive Learning Approach for Anomaly Detection in Multi-View Scenarios

Paula Mielgo[1], Anibal Bregon[2], Carlos J. Alonso-González[3], Miguel A. Martínez-Prieto[4] and Belarmino Pulido[5]

[1,2,3,4,5] *Department of Computer Science, University of Valladolid, Valladolid, Spain*
paula.mielgo@uva.es, anibal.bregon@uva.es, calonso@uva.es, miguelamp@uva.es, b.pulido@uva.es

ABSTRACT

Quality control is a key task in smart manufacturing, since it ensures that processes consistently meet rigorous performance standards. The effective implementation of these mechanisms is crucial to ensuring both reliability and efficiency in modern manufacturing environments, where automation is increasingly integrated. Traditional anomaly detection algorithms typically rely on single-view data for each manufacturing product, overlooking relevant and complementary information available from multiple perspectives. Furthermore, cross-entropy-based loss functions are frequently adopted in the literature to train detection models; however, these approaches often struggle with imbalanced datasets or when detecting rare and subtle anomalies. In this work, a contrastive learning architecture for multi-view anomaly detection in industrial settings is proposed. The method performs a mid-level fusion to generate a structured representation of the input instances, thereby enhancing detection capabilities. The architecture was evaluated on the Real-IAD dataset, where it demonstrated better performance than traditional techniques. These findings highlight the potential of contrastive learning to improve anomaly detection performance, thus contributing to the construction of more reliable quality control systems in smart manufacturing environments.

1. INTRODUCTION

In the context of Industry 4.0, modern manufacturing systems are becoming increasingly reliant on real-time data acquisition and processing. As a result, the ability to automatically detect deviations from expected behavior is critical for ensuring product quality and optimizing process efficiency. The effective identification of faults, errors, and abnormal conditions plays a key role in maintaining system reliability and robustness. Consequently, anomaly detection represents a fundamental task in industrial environments. Given the increasing complexity of modern industrial systems, it becomes essential to collect sufficient data to support accurate decision-making. For instance, when detecting defects in a manufactured product, capturing images from a single viewpoint will not reveal all possible anomalies if the defect is located on a different perspective. This encourages the value of the multi-view approach, which consists of the utilization of data collected from multiple perspectives of the same instance. Multi-view data provides a more comprehensive representation of the object being monitored, allowing for more complete and consistent analysis. For these reasons, multi-view data is becoming increasingly common in industrial applications, better reflecting real-world conditions.

Multi-view problems have typically been addressed through different approaches (S. Wang et al., 2022), aiming to exploit the complementary information present across different views to enhance learning performance. Among these, fusion-based methods have proven particularly effective, achieving strong results by integrating multiple representations into a unified embedding. In particular, such methods have been successfully employed in conjunction with Convolutional Neural Network (CNN) architectures (Khan, Shahid, Raza, Dar, & Alquhayz, 2019) and attention mechanisms (He, Zhang, Tian, Wang, & Xie, 2024). Despite their efficiency, multi-view fusion-based methods pose several challenges (Yu et al., 2025), including inconsistent and often imbalanced information across views. For instance, some views may include irrelevant information or even be redundant with others. Additionally, effectively combining different views into a single and meaningful joint representation increases the overall complexity, both in terms of structure and scalability.

When working with visual data, such as images or videos, structuring the representation space effectively due to the high dimensionality of the input data becomes critical. To address this, contrastive learning has emerged as a powerful Machine Learning (ML) tool for learning data representations by pulling together similar (positive) pairs and pushing apart dissimilar (negative) pairs. To do it, it employs a loss function that minimizes the distance between positive pairs

while maximizing the distance between negative pairs. This approach has been widely applied to computer vision tasks for both self-supervised and supervised learning purposes. In (Chen, Kornblith, Norouzi, & Hinton, 2020), the proposed SimCLR architecture uses data augmentation to construct image pairs from unlabeled data. The model, which consists of an encoder and a nonlinear projection head, outperforms previous methods by introducing the NT-Xent loss function. A similar architecture adapted for supervised learning tasks is used in (Khosla et al., 2020). In this case, the objective is to reduce the distance between elements of the same class using a contrastive loss function. This method achieves better results than traditional cross-entropy approaches in large-scale classification problems. Contrastive learning has also been applied to anomaly detection. In (Kopuklu, Zheng, Xu, & Rigoll, 2021), a contrastive learning framework is proposed to identify abnormal driving behaviors. This approach demonstrates superior performance to previous methods and also improves model robustness.

A novel contrastive learning architecture for anomaly detection in multi-view environments is proposed in this work. Unlike traditional approaches that process each view independently or perform a late fusion of them, this solution employs mid-level fusion, allowing for the effective integration of information from multiple views. The architecture also leverages the properties of contrastive learning to structure the latent space in a way in which normal instances are grouped together and the defective ones are clearly separated. The main contributions of this work are: (1) a novel architecture based on contrastive learning to effectively handle multi-view data; and (2) its implementation and evaluation for anomaly detection in an industrial environment. To the best of our knowledge, this work represents the first application of contrastive learning for anomaly detection on an industrial multi-view dataset. The proposed architecture demonstrated strong performance in industrial anomaly detection, achieving an average AUC-ROC of 0.986 and a perfect score of 1.000 in four categories. Compared to the baseline, it outperformed in 7 out of 13 categories and showed lower performance in only 2.

The remaining of the paper is organized as follows. Section 2 introduces the multi-view approach for quality control purposes, along with the contrastive learning technique. Section 3 presents the proposed architecture for anomaly detection in multi-view datasets. Section 4 describes the selected dataset, the experimental setup, and the results for both the baseline and the proposed method. Finally, Section 5 summarizes the main conclusions and outlines future directions for this work.

2. BACKGROUND

This section presents different multi-view approaches for image anomaly detection and introduces contrastive learning.

2.1. Multi-View Approaches for Visual Anomaly Detection

Anomaly detection is a fundamental challenge in modern manufacturing processes. Traditionally, this task has been addressed through the analysis of numerical data obtained from sensors. Nevertheless, the recent advent of Deep Learning (DL) has enabled the employment of images captured during the manufacturing process to identify defective parts. The literature on visual anomaly detection has mainly centered on single-view problems, where datasets consist of one image per product. Common DL approaches to address these tasks include embedding-based methods (Defard, Setkov, Loesch, & Audigier, 2021), which are focused on learning a latent representation of data; reconstruction-based methods (Park, Lee, Ko, & Kim, 2023), which enhance the training set with transformed or synthetic samples to improve model performance. However, the single-view approach presents challenges, as even when an object has a visible anomaly, it might be located on the opposite side that the camera cannot capture. Therefore, to better reflect real-world conditions, the multi-view problem was introduced several years ago. Multi-view datasets consist of image sets captured from different angles of each product in order to detect all possible defects. They provide an effective method for representing 3D objects through 2D image collections while preserving most of the relevant information.

As stated in (S. Wang et al., 2022), multi-view approaches can be classified into four distinct categories. First, fusion-based solutions, which are designed to combine the representations from each view into a joint embedding. This unified representation can be then processed by a single-view anomaly detection algorithm. Second, alignment-based solutions, which enforce consistency among the representations learned from different views by aligning them into the latent space. That is, representations obtained from each view are trained to be similar to each other. Third, deep anomaly detection tailored solutions, which are focused on training a deep anomaly detection model for data from each view and then combining each individual output to construct the final result. Finally, self-supervision-based approaches design pretext tasks, allowing the model to learn meaningful representations without external supervision. These tasks exploit the inherent structure of multi-view data to guide training, and the reconstruction error can serve as an anomaly score for the final prediction. The work in (S. Wang et al., 2022) provides an overview of representative methods across all four categories. Fusion-based approaches have received particular attention, as recent studies have demonstrated their strong performance and effectiveness in integrating complementary information from multiple views. In (Su, Maji, Kalogerakis, & Learned-Miller, 2015) a fusion-based solution that outperforms methods operating directly on 3D shape representations is presented. The developed architecture, based on CNNs, employs

a VGG-M model that has been pretrained on ImageNet. This model is then fine-tuned to extract features from each view, followed by a view-pooling layer. Then, a second CNN processes the fused representation to perform the final classification. In (He et al., 2024) the authors propose a multi-view anomaly detection (MVAD) framework that uses attention mechanisms to efficiently combine information from different views. The framework learns a fused representation of the different views using the Multi-View Adaptive Selection (MVAS) algorithm, which selects and integrates the most relevant local regions across views. This representation is then used to detect anomalies from a single embedding. In (Jakob, Madan, Schmid-Schirling, & Valada, 2021) the authors introduce the Dices dataset, consisting of 2000 grayscale images of falling dice captured from multiple perspectives, with 5% of the data containing anomalies. They also propose a Deep Support Vector Data Description (Deep SVDD) algorithm, employing different fusion techniques. Additionally, data augmentation and denoising methods are applied to enhance the robustness of the model in the presence of noise. Evaluations conducted on both the Dices dataset and MNIST demonstrate that their proposed architecture achieves better results than single-view anomaly detection methods.

2.2. Contrastive Learning

Contrastive learning is a ML technique that constructs models to distinguish between similar and dissimilar pairs of examples, which are denoted as positive and negative pairs, respectively. These models learn useful representations of data in an embedding space by grouping positive pairs together and separating negative ones. To achieve this, a contrastive loss function is employed to maximize the distance between negative pairs while minimizing the distance between positive pairs. Therefore, it is essential to correctly define what constitutes a positive or a negative pair. Commonly, positive pairs consist of elements belonging to the same entity or class, while negative pairs consist of elements from different classes. It is important to note that labeled data is often not strictly required for this approach.

Visual data usually exhibit more complex latent representations due to their high dimensionality. Consequently, contrastive learning represents an effective approach to structuring the embedding space. One of the first applications of a contrastive learning loss function is (Hadsell, Chopra, & LeCun, 2006), which employs a siamese architecture and defines positive and negative pairs based on neighborhood relationships. Experiments demonstrate that the contrastive loss function maintains an equilibrium in the output space and prevents the system from collapsing into a constant function. (Schroff, Kalenichenko, & Philbin, 2015) presented the triplet loss. For each image of a specific person, denoted as the anchor, a triplet is formed by selecting positive elements, which are other images from the same person, and negative elements, with images from different people. The proposal also incorporates an α parameter to ensure a minimum separation between the components. Moreover, to improve the process and optimize its efficiency, a triplet selection process is implemented. This method has been demonstrated to simplify the setup process and enhance performance. (Chen et al., 2020) introduces SimCLR, a self-supervised learning method that learns visual representations by contrasting positive pairs with negative pairs. Since positive pairs are constructed by the application of data augmentation to original data, labeled data is not required. The architecture is composed of three principal components: an encoder, a nonlinear projection head, and the NT-Xent loss function, introduced in the same work. The proposal outperforms previous methods for self-supervised, semi-supervised and transfer learning, while simplifying existing architectural approaches. (Khosla et al., 2020) presents an extension of the contrastive learning approach to supervised problems, leveraging label information to improve learned representations. Thus, the idea is to not only try to set the augmented data closer to the original, but also setting elements from the same class closer, structuring the embedding space. The architecture is similar to the SimCLR, combining an encoder network with a projection head. Moreover, the work compares two different versions of the supervised contrastive loss, outperforming traditional approaches that employ cross-entropy loss on large-scale classification problems. Anomaly detection is another area where these methods are applied. (Kopuklu et al., 2021) introduces a novel dataset specifically for driver anomaly detection. This work also proposes a contrastive learning approach to effectively identify unusual driving behaviors by learning to distinguish between normal and anomalous patterns. It demonstrates that this method significantly improves the accuracy and robustness of detecting various driver anomalies in comparison with the cross-entropy loss.

3. ARCHITECTURE PROPOSAL

As previously stated, real-world scenarios are better addressed using multi-view approaches. Additionally, the benefits of contrastive learning methods were discussed. Therefore, the proposed architecture leverages the power of contrastive learning methods for anomaly detection in multi-view settings. The method is classified in the fusion-based category, as it performs a mid-level fusion of features extracted from each view to finally generate a unified output. An overview of the proposed architecture can be consulted in Figure 1. The architectural framework is composed of two main components:

- Base encoder, which serves as a feature extractor. It is composed of one CNN per view. Each CNN transforms an input image $i \in \mathbb{R}^{H \times W \times C}$, where H is the height, W the width, and C the number of channels, into a feature

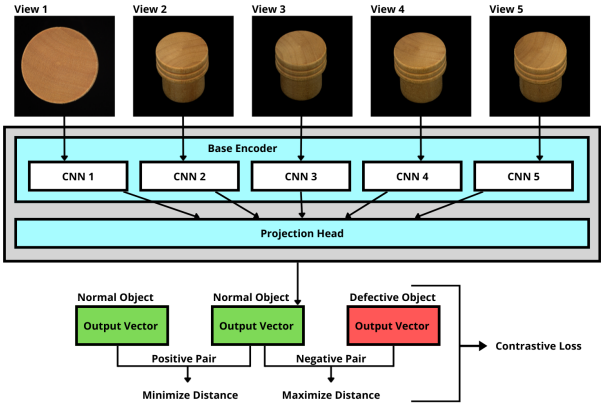

Figure 1. Architecture proposal.

vector, $v \in \mathbb{R}^M$, where M is the output dimension of the CNN.

- Projection head, which maps the concatenated feature vectors to a joint representation space. This component consists of a sequence of linear layers. The outputs of the CNNs are $v_1, v_2, ..., v_n \in \mathbb{R}^M$, where n is the number of views. These vectors are concatenated and passed through the projection head to produce a single embedding vector $z \in \mathbb{R}^L$, where L is the dimensionality of the latent space.

For the training process a contrastive loss function is employed. In particular, the loss function used is the one presented in (Kopuklu et al., 2021), as it is tailored for anomaly detection tasks. This function is designed to bring normal instances closer together in the latent space, while pushing defective instances away from them. Thus, positive pairs are formed exclusively from the normal (non-defective) class, while negative pairs consist of one element from the normal class and one from the defective class. Each batch contains a proportional number of elements from both classes, according to their distribution in the dataset. Let N be the number of normal instances in a batch, denoted as $n_r,\ r \in \{1, 2, ..., N\}$, and D the number of defective instances, denoted as $d_s,\ s \in \{1, 2, ..., D\}$. Therefore, there are $N(N-1)$ positive pairs, and ND negative pairs in this batch. The contrastive loss function is detailed in Eq. 2.

$$\mathcal{L}_{i,j} = -\log \frac{\exp(sim(n_i, n_j)/\tau)}{\exp(sim(n_i, n_j)/\tau) + \sum_{k=1}^{K} \exp(sim(n_i, d_k)/\tau)} \tag{1}$$

$$\mathcal{L} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{1}_{i \neq j} \mathcal{L}_{i,j} \tag{2}$$

where $sim(\cdot, \cdot)$ is the cosine similarity function between two vectors, τ is a positive temperature parameter, and $\mathbb{1}$ is the indicator function.

4. RESULTS

This section begins by introducing the dataset and the experimental setup. Thereafter, results of both the baseline and the proposed method are discussed.

4.1. Dataset

The dataset selected to perform the architecture evaluation is the Real-IAD dataset (C. Wang et al., 2024). The Real-IAD dataset is a realistic, large-scale, and multi-view collection for industrial anomaly detection, that contains 150000 images of 30 different objects. It includes high-resolution images and various defect types, with a larger range of defect proportion. Compared to other industrial datasets such as the MVTec AD (Bergmann, Fauser, Sattlegger, & Steger, 2019), which has achieved over 99% in AUC-ROC using state-of-the-art methods, Real-IAD provides a more challenging benchmark that encourages the development of novel and more robust solutions. It is the first dataset for anomaly detection that presents a multi-view setting, with each instance captured from five different angles. Furthermore, since approximately one-third of the instances belong to the defective class, the dataset also enables supervised learning tasks.

The dataset distinguishes between two types of object categories: those that are symmetrical in four of the five views, and those that show distinct from all five views. Experiments were performed on objects in the first category. A visual overview of selected objects from the dataset is provided in Figure 2.

4.2. Experimental Setup

Experiments were conducted on an Intel Xeon Silver 4310 CPU (2.10 GHz) with 32 GB of RAM. Moreover, an NVIDIA A40 GPU with 48 GB of memory was used for accelerated computation. The environment was configured with Python 3.9.12, PyTorch 1.13.1 and Scikit Learn 1.0.2 as the main DL frameworks.

For the base encoder, the ResNet-18 architecture was selected. Specifically, a pretrained model from TorchVision, trained on ImageNet, was chosen to be subsequently fine-tuned. Each CNN model produces a 512-dimensional feature representation. The projection head is composed of a sequence of two fully connected layers, with a ReLU activation function applied between them. The output dimensionality of this component is set to 32. For fair comparison, the same pretrained ResNet-18 model was also used in the baseline.

Images were resized to 224x224 and normalized using a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225] for the red, green, and blue channels, respectively, following the preprocessing scheme employed in the pretrained model. The batch size was set to 16, the learning rate to 0.0001, and the τ parameter to 0.1. For both the base-

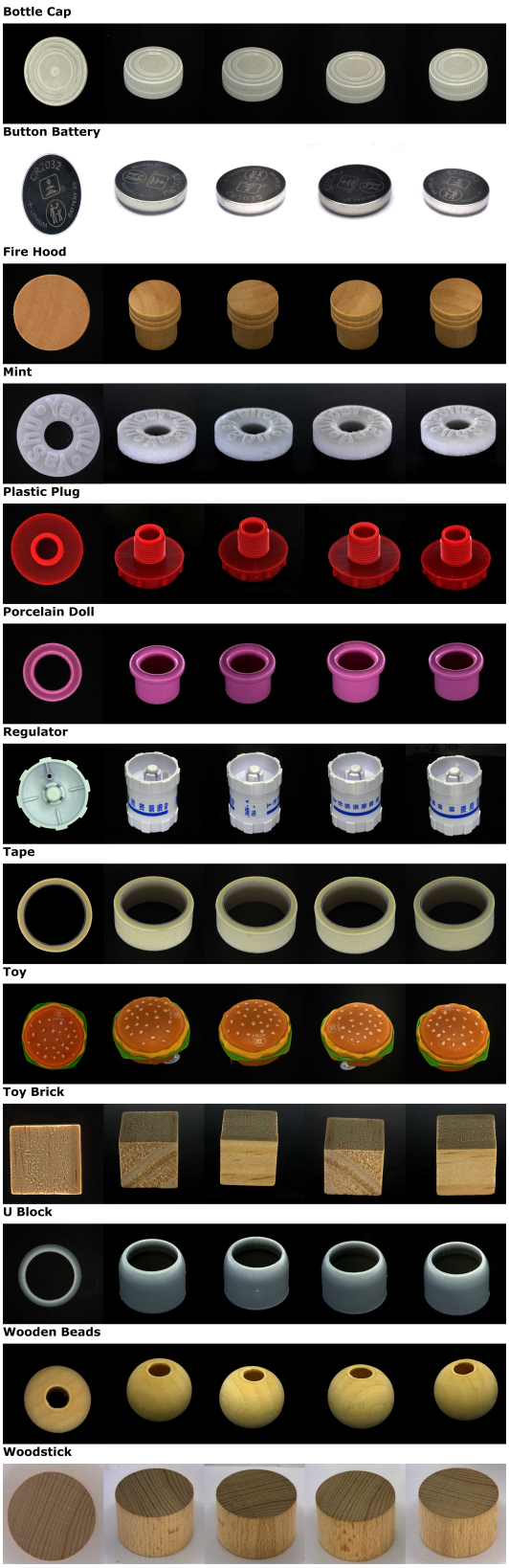

Figure 2. Real-IAD dataset.

line and the proposed model training processes, the Adam optimization algorithm was selected and an early stopping criterion with a patience of 20 was applied based on the validation loss during 1000 epochs.

The architectures were evaluated using metrics derived from the components of the confusion matrix. In binary classification, they includes:

- True positive (TP). The actual and predicted values are both positive.
- False positives (FP). The actual value is negative, while the predicted is positive.
- False negatives (FN). The actual value is positive, while the predicted is negative.
- True negative (TN). The actual and predicted values are both negative.

In this work, the defective class is considered the positive one. The selected evaluation metric is the area under the curve ROC (AUC-ROC). In binary classification tasks, the model outputs a probability score representing the likelihood that an instance belongs to the positive class. A final classification decision is then made by applying a threshold (typically set at 0.5) to this score. The ROC curve provides a visual representation of the performance of the model across different threshold values. It is generated by discretizing the interval [0,1] and evaluating the True Positive Rate (TPR) (Eq. 3) and the False Positive Rate (FPR) (Eq. 4) at each point.

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

The ROC curve is then plotted by representing TPR against FPR for each threshold. Finally, the AUC-ROC is obtained by integrating the ROC curve with respect to the FPR axis.The AUC-ROC metric represents the probability that the model assigns a higher confidence score to a randomly selected positive instance than to a negative one. This means that:

- If AUC-ROC > 0.5, the model performs better than a random classifier.
- If AUC-ROC ≤ 0.5, the model performs no better than a random classifier.

AUC-ROC is an effective method for evaluating classifiers, as it offers a threshold-independent evaluation, balances sensitivity and specificity, and provides an intuitive interpretation of model performance.

4.3. Baseline Results

Two baselines are established to compare the proposed architecture. First, one classifier per view is trained to perform individual classifications. During inference, each view of the instance is passed through its corresponding classifier, generating a partial decision based on the individual probability score with a threshold of 0.5. The individual predictions are then combined to generate the final decision as follows:

- If any of the individual predictions is defective, the final decision is defective.
- Otherwise, the final decision is normal.

The results obtained using this approach are presented in Table 1. It is important to note that, since the final prediction is made using a logical OR rule, the complete architecture does not produce a probabilistic score that can be used to compute the AUC-ROC metric.

Table 1. Classification test results for the first baseline architecture.

Category	TP	FP	FN	TN
Bottle Cap	103	11	1	88
Button Battery	102	4	2	93
Fire Hood	103	40	0	61
Mint	144	60	3	7
Plastic Plug	104	38	0	59
Porcelain Doll	100	9	5	88
Regulator	105	38	0	62
Tape	104	31	0	69
Toy	102	21	1	79
Toy Brick	101	45	2	55
U Block	101	20	0	80
Wooden Beads	108	32	7	55
Woodstick	101	43	0	58

The number of false positive instances is considerably high across all categories, with the Mint category reaching a false positive rate of 0.9. Therefore, a more realistic baseline is introduced to better demonstrate the effectiveness of the contrastive loss. To this end, an architecture similar to the one proposed in Section 3 is constructed. The main difference is that, in this baseline, Binary Cross-Entropy is used as the loss function. Furthermore, to ensure compatibility with this loss function, the output dimensionality of the projection head is modified to 1. Once again, classification is performed using a threshold of 0.5.The corresponding results are detailed in Table 2.

This approach has been demonstrated to achieve competitive metrics, reducing the number of normal elements misclassified as defective. The AUC-ROC score exceeds 0.881 in all categories, reaching 1.000 in four of them. However, the Mint category still shows poor performance, suggesting the challenging nature of this category.

Table 2. Classification test results for the second baseline architecture.

Category	TP	FP	FN	TN	AUC-ROC
Bottle Cap	102	0	2	99	0.990
Button Battery	102	0	2	97	0.986
Fire Hood	103	0	0	101	1.000
Mint	116	10	31	57	0.881
Plastic Plug	103	1	1	96	0.992
Porcelain Doll	100	0	5	97	0.983
Regulator	104	0	1	100	0.993
Tape	104	0	0	100	1.000
Toy	102	0	1	100	0.998
Toy Brick	100	3	3	97	0.982
U Block	101	0	0	100	1.000
Wooden Beads	104	3	11	84	0.967
Woodstick	101	1	0	100	1.000

4.4. Proposal Results

Experimental results obtained using the proposed architecture are discussed in this section. Since the proposed method learns a representation of each object in a latent space, bringing normal instances closer together while separating the defective ones, a visual representation of the results for all categories is presented in Figure 3. This visualization uses the t-SNE algorithm (Maaten & Hinton, 2008) to represent the output vectors into two dimensions.

The proposed approach appears effective in achieving its objective. To enable a more direct comparison with the baseline, a classification head is appended after the projection head. Logistic Regression is selected as the classification method, with the decision threshold set to 0.5. The corresponding results are presented in Table 3.

Table 3. Classification test results for the proposed architecture.

Category	TP	FP	FN	TN	AUC-ROC
Bottle Cap	102	0	2	99	0.995
Button Battery	102	0	2	97	0.993
Fire Hood	103	0	0	101	1.000
Mint	121	8	26	59	0.896
Plastic Plug	103	3	1	94	0.999
Porcelain Doll	100	0	5	97	0.979
Regulator	104	0	1	100	0.999
Tape	104	0	0	100	1.000
Toy	102	0	1	100	0.997
Toy Brick	100	0	3	100	0.983
U Block	101	0	0	100	1.000
Wooden Beads	103	3	12	84	0.974
Woodstick	101	0	0	101	1.000

The proposal results are also competitive. The AUC-ROC score has a minimum value of 0.896 and reaches 1.000 in four categories. In the Mint category, performance is slightly improved. Since the results may appear similar to those of the baseline, a direct comparison is presented in Table 4.

The comparison reveals a generally better performance of

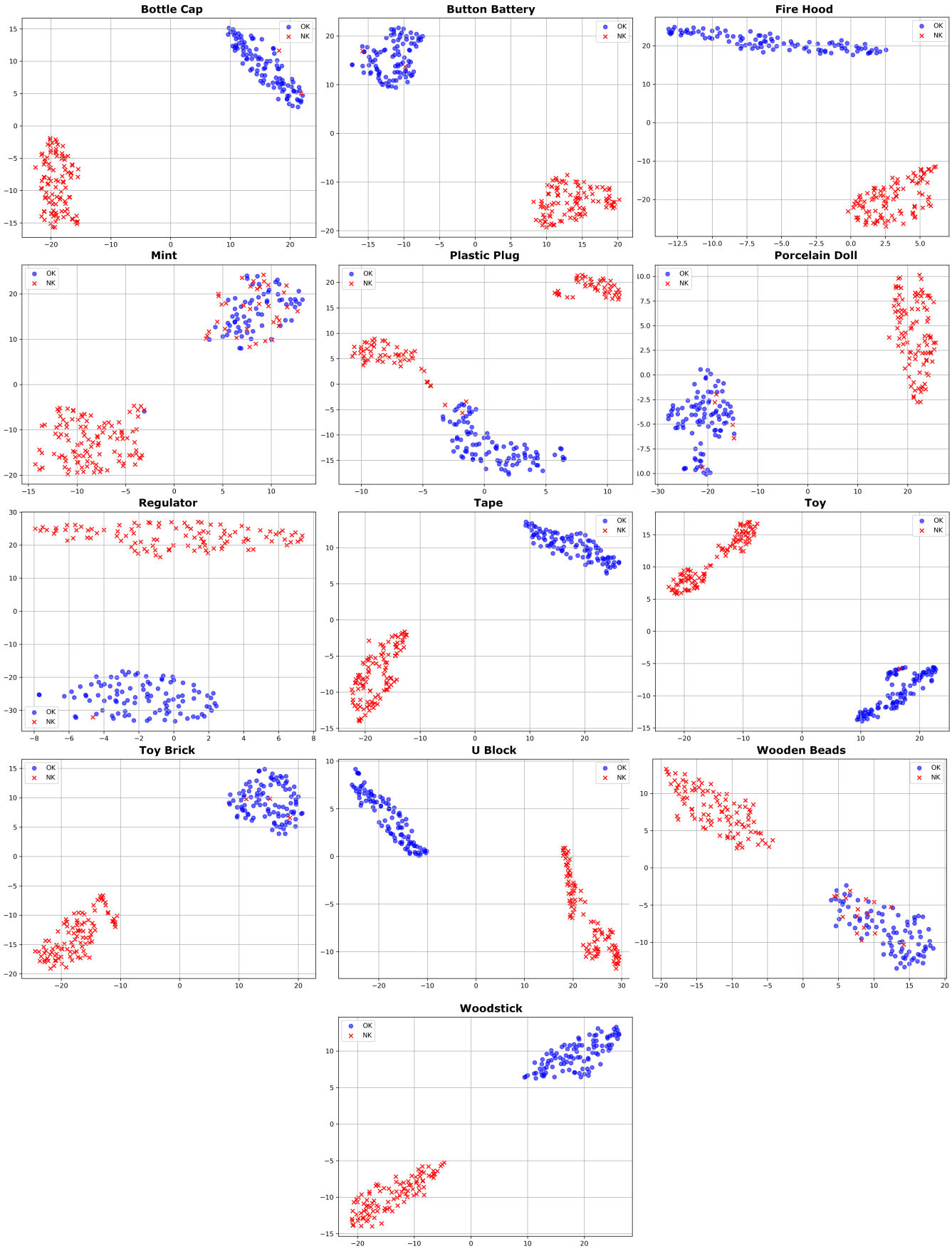

Figure 3. Outputs representation.

Table 4. AUC-ROC comparison between baseline and proposed approach.

Category	Baseline AUC-ROC	Proposal AUC-ROC
Bottle Cap	0.990	**0.995**
Button Battery	0.986	**0.993**
Fire Hood	**1.000**	**1.000**
Mint	0.881	**0.896**
Plastic Plug	0.992	**0.999**
Porcelain Doll	**0.983**	0.979
Regulator	0.993	**0.999**
Tape	**1.000**	**1.000**
Toy	**0.998**	0.997
Toy Brick	0.982	**0.983**
U Block	**1.000**	**1.000**
Wooden Beads	0.967	**0.974**
Woodstick	**1.000**	**1.000**
Mean	0.982	**0.986**

the proposed architecture, as evidenced by the superior mean value. Analyzing the categories individually, the proposed method outperforms in 7 out of the 13 categories. Moreover, it achieves identical results in 4 categories, with an AUC-ROC value of 1.000, which represents the maximum possible score. The proposed approach shows inferior performance in only 2 categories.

5. CONCLUSIONS AND FUTURE WORK

In this work, a novel architecture combining a fusion-based approach with a contrastive loss function was proposed. The method has proven to be effective in visual anomaly detection scenarios. It combines a base encoder and a projection head to structure the latent space, effectively separating normal instances from defective ones. Additionally, this approach has demonstrated superior classification performance compared to the traditional Binary Cross-Entropy loss function.

The proposed method achieved a minimum AUC-ROC of 0.896, reaching a perfect score of 1.000 in four categories. It performed worse than the second baseline in only two categories, obtaining equal or better results in the remaining eleven. Furthermore, the average value obtained using the proposal is higher than the baseline.

The future work will focus on extending this method to the remaining categories of the Real-IAD dataset. Additionally, future work will include the study of the computational complexity of the contrastive learning loss function and the exploration of potential modifications to reduce it. Finally, we plan to analyze the generalization capacity of the model to unknown defects by excluding some defective categories from the training set. This consideration is crucial since, in practical applications, it is not feasible to obtain samples for all possible types of defects, and the model should still provide reliable performance under these circumstances.

ACKNOWLEDGMENT

Paula Mielgo's work has been funded by UVa 2023 predoctoral contracts, co-financed by Banco Santander. This work has been funded by the Spanish Ministerio de Ciencia e Innovación under grant PID2021-126659OB-I00.

REFERENCES

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9592–9600).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Defard, T., Setkov, A., Loesch, A., & Audigier, R. (2021). Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition* (pp. 475–489).

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition* (Vol. 2, pp. 1735–1742).

He, H., Zhang, J., Tian, G., Wang, C., & Xie, L. (2024). Learning multi-view anomaly detection. *arXiv preprint arXiv:2407.11935*.

Jakob, P., Madan, M., Schmid-Schirling, T., & Valada, A. (2021). Multi-perspective anomaly detection. *Sensors*, *21*(16), 5311.

Khan, H. N., Shahid, A. R., Raza, B., Dar, A. H., & Alquhayz, H. (2019). Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access*, *7*, 165724–165733.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, *33*, 18661–18673.

Kopuklu, O., Zheng, J., Xu, H., & Rigoll, G. (2021). Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 91–100).

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Park, S., Lee, K. H., Ko, B., & Kim, N. (2023). Unsupervised anomaly detection with generative adversarial networks in mammography. *Scientific Reports*, *13*(1), 2925.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet:

A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3D shape recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 945–953).

Wang, C., Zhu, W., Gao, B.-B., Gan, Z., Zhang, J., Gu, Z., … Ma, L. (2024). Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22883–22892).

Wang, S., Liu, J., Yu, G., Liu, X., Zhou, S., Zhu, E., … Yang, W. (2022). Multiview deep anomaly detection: A systematic exploration. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(2), 1651–1665.

Yu, Z., Dong, Z., Yu, C., Yang, K., Fan, Z., & Chen, C. P. (2025). A review on multi-view learning. *Frontiers of Computer Science*, *19*(7), 197334.

BIOGRAPHIES

Paula Mielgo is a Ph.D. Student in the Computer Science Department at the University of Valladolid, Spain. Mielgo completed her two Bachelor's Degrees in Mathematics and Computer Science Engineering in 2022. Furthermore, she obtained a Master's Degree in Computer Science Engineering in 2023. During her studies, she received several academic awards, including the award for the best academic results in both of her CS programs and a regional award for her final degree thesis. Later that year, she started her Ph.D. thesis in Computer Science.

Anibal Bregon received his B.Sc., M.Sc. and Ph.D. degrees in Computer Science from the University of Valladolid (Spain) in 2005, 2007 and 2010, respectively. He joined the Department of Computer Science at the University of Valladolid in 2011, where he is Associate Professor since February 2018. He has carried out both basic and applied research in the areas of fault diagnosis and prognosis for aerospace and industrial systems, and has co-authored more than 85 journal and conference papers. He is currently leading a national funded project on advanced learning for smart manufacturing and several technology transfer contracts on Deep Learning, and has also participated as researcher on several funded projects, networks and contracts on fault diagnosis and prognosis topics, on Big Data analytics and on Deep Learning. He has been guest researcher with the Intelligent Systems Division at NASA Ames Research Center and the Institute for Software Integrated Systems at Vanderbilt University, among others. His current research interests include model-based reasoning for diagnosis and prognosis, health-management, Big Data, Industry 4.0 and Deep Learning. Among various other professional activities, he has held different chair positions at the PHM and PHME conferences, has been co-administrator of several courses and summer schools on diagnosis, prognosis, and artificial intelligence, and has been the Local Chair of the 2016 European Conference of the Prognostics and Health Management Society.

Carlos J. Alonso-González received the B.S. and Ph.D. degrees in physics from the University of Valladolid, in 1985 and 1990, respectively. After a brief stay in private companies and the Public University of Navarra, he joined the University of Valladolid, where he is currently an Associate Professor with the Department of Computer Science. He is also the Head of the Intelligent Systems Group, Department of Computer Science. He has worked on different national and European-funded projects related to the monitoring and diagnosis of continuous industrial environments and dynamic hybrid systems. He is also involved in projects related to the application of deep learning and casual and explainable AI to Industry 4.0, both for manufacturing and continuous processes. His current research interests include knowledge, model, and data-based systems for health management of dynamic systems, model and data-based diagnosis and prognosis of complex physical systems, and machine learning. He has been a member of the Board of Trustees of the Sugar Technology Center, being responsible for projects related to the application of artificial intelligence to online production supervision.

Miguel A. Martínez-Prieto is an Associate Professor and Researcher in Computer Science with the Department of Computer Science, University of Valladolid, Spain. He received his B.Sc., M.Sc., and Ph.D. degrees in Computer Science from the University of Valladolid, Valladolid, Spain, in 2005, 2007, and 2010, respectively. He held a Postdoctoral position with the Department of Computer Science, University of Chile, from 2010 to 2012. His research has been in the area of data management, mainly in data compression and indexing of semantic, text and biological data, and the resolution of specific queries in each of these scenarios. His current research interests focus on data science, with applications in air traffic management and Industry 4.0. He has co-authored more than 90 peer-reviewed papers on these topics and has been involved in several European and national funded projects, as well as several transfer contracts with companies and government institutions.

Belarmino Pulido received his Licenciate degree, M.Sc. degree, and Ph.D. degree in Computer Science from the University of Valladolid, Valladolid, Spain, in 1992, 1995, and 2001 respectively. In 1994 he joined the Department of Computer Science at the University of Valladolid, where he is Associate Professor since 2002. His main research interests are in Systems Health Management using different techniques such as model-based reasoning, knowledge-based systems, and data-driven models (using both classical and advanced machine-learning). He has worked in different national and european funded projects related to Supervision and Diagnosis. He is the coordinator of the Spanish Network on Supervision and Diagnosis of Complex Systems since 2005.