

Towards Sim-to-Real Industrial Parts Classification with Synthetic Dataset

¹Yanna Prasanth, ²Dr. M. Kavitha

^{1,2} Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India,

Abstract:- Automated industrial part classification is important for intelligent manufacturing, robotic inspection, and automated quality control, where accurate visual recognition supports efficient and consistent component identification. However, conventional approaches often require large collections of real-world labeled images, while variations in illumination, background, orientation, and image quality can adversely affect model generalization. The Synthetic Industrial Parts Dataset (SIP-17) is utilized as the source dataset, with the Synthetic data with background subset used for model training and validation and the Realdata test subset reserved for real-world testing. Images are resized and augmented using random horizontal and vertical flipping, rotation, color jittering, tensor conversion, and normalization. ResNet50, ConvNeXt Tiny, EfficientNet-B0, EfficientNet-B2, Swin Transformer Tiny, ViT-B/16, MobileNetV3, CondConv-EfficientNet, MaxViT-Tiny, and Xception are comparatively evaluated. Model performance is assessed using accuracy, precision, recall, and F1-score on validation and real-world test data to examine sim-to-real generalization. Among the evaluated architectures, MaxViT-Tiny achieves the best performance, attaining 99.8% validation accuracy and 91% accuracy on the real-world test dataset. The resulting system integrates FastAPI-based inference with Grad-CAM++ explanations, providing practical and interpretable industrial part classification under real-world domain variation.

Keywords: Sim-to-real learning, industrial parts classification, synthetic data, domain generalization, deep learning, Grad-CAM++.

1. Introduction

Industrial automation increasingly depends on computer vision for recognizing and understanding physical components in manufacturing and robotic environments. Reliable identification of industrial parts can support inspection, assembly, sorting, maintenance, and automated handling, where incorrect recognition may affect subsequent operations. Conventional vision methods often depend on manually designed visual characteristics and controlled imaging conditions, which can limit their adaptability to complex industrial environments. Deep learning has provided a more flexible alternative by enabling visual representations to be learned directly from image data. Its application to industrial part identification has demonstrated the potential to recognize components based on their inherent visual characteristics rather than relying exclusively on manually defined features [14]. The increasing availability of advanced visual learning architectures has further strengthened the role of data-driven approaches in industrial computer vision.

A major challenge in developing such systems is the availability of sufficient real-world training data. Industrial components can occur in different poses, backgrounds, illumination conditions, and visual configurations, making comprehensive data collection difficult. Creating accurate annotations for large collections of industrial images can also require substantial human effort. Computer-aided design and synthetic image generation offer an alternative by enabling visual data to be produced from digital representations of objects. CAD-based learning has been investigated for industrial object recognition [1], while synthetic dataset generation has demonstrated potential for object-to-model learning in industrial applications [19]. These approaches provide greater control

over the creation of training examples and can reduce the dependence on extensive manually collected real-world datasets.

However, training with synthetic imagery introduces an important challenge known as the simulation-to-real gap. Synthetic images may differ from real images in their appearance, environmental context, textures, lighting, and other visual characteristics. Consequently, a model that performs effectively on synthetic images may experience reduced performance when exposed to real-world observations. Domain randomization has been proposed as a way of increasing the diversity of simulated environments and improving transfer to real settings [17]. Research on structured domain randomization has further emphasized the importance of controlling contextual variation when generating synthetic training data [13]. Although these approaches demonstrate the value of synthetic data for transfer learning, achieving reliable generalization remains challenging because the visual differences between generated and real environments cannot always be fully represented during training.

The development of increasingly capable deep learning architectures provides further opportunities for addressing visual recognition challenges. Residual learning established an effective approach for training deep convolutional networks [6], while EfficientNet demonstrated the importance of balancing model capacity and computational efficiency through systematic scaling [15]. Vision Transformers introduced an alternative representation-learning paradigm by treating images as sequences of visual patches [4]. More recent convolutional architectures have also demonstrated that modern network designs can achieve strong visual recognition performance [9]. Despite these advances, the effectiveness of different architectures can vary considerably when the training and evaluation domains differ. Existing approaches therefore leave a need for more systematic investigation of model suitability specifically under synthetic-to-real industrial recognition conditions.

This motivates the development of a unified approach for reliable industrial part classification across simulated and real visual domains. The objective is to investigate how effectively visual models trained using synthetic industrial imagery can generalize to real-world imagery while providing meaningful and understandable predictions. The proposed approach focuses on the relationship between synthetic-data-based learning, model selection, real-world generalization, and explainable visual prediction. By bringing these aspects together, the approach aims to reduce dependence on large-scale real-world training data while improving confidence in automated industrial recognition. The resulting framework addresses the broader need for scalable and interpretable computer vision systems capable of supporting industrial environments where data availability and domain variation remain significant challenges.

2. Related Work

Research in industrial computer vision has increasingly explored deep learning for automated recognition and identification of physical components. Krueger et al. investigated deep learning for part identification based on inherent visual features and demonstrated the potential of learned representations for recognizing industrial parts [14]. Cohen et al. examined CAD-based learning for object detection in an industrial context, showing how digital design information can support visual learning when real-world data is limited [1]. Similarly, Mayershofer et al. introduced the LOCO dataset to support recognition of logistics-related objects in realistic environments [10]. These studies established important foundations for applying data-driven visual recognition to industrial and logistics scenarios. However, acquiring sufficiently diverse real-world imagery remains a significant challenge when objects must be recognized across varying environments and operating conditions.

Synthetic data has consequently received considerable attention as an alternative source of training imagery. Peng et al. investigated learning object detectors from 3D models, demonstrating that rendered object representations can contribute to training visual recognition systems [11]. Richter et al. explored the use of computer-generated environments for obtaining training data and showed the potential of virtual scenes for large-scale visual learning [14]. Wong et al. focused specifically on synthetic dataset generation for industrial applications and demonstrated how generated imagery can support object-to-model learning [19]. Eversberg and Lambrecht further investigated physics-based rendering for industrial object detection and compared visual realism with domain randomization [5]. These contributions indicate that synthetic imagery can reduce the burden associated with collecting and

annotating large real-world datasets. Nevertheless, differences between generated and real imagery can still influence recognition performance.

The challenge of transferring knowledge from simulated environments to real environments has been widely examined through domain randomization. Tobin et al. introduced domain randomization as a strategy for transferring models trained in simulation to real-world settings by exposing models to diverse simulated appearances [17]. Tremblay et al. investigated synthetic training combined with domain randomization and demonstrated its usefulness for reducing the reality gap [18]. Prakash et al. extended this concept through structured domain randomization, emphasizing contextual variation when generating synthetic training data [13]. Yue et al. investigated domain randomization together with pyramid consistency to improve simulation-to-real generalization without requiring access to target-domain training data [20]. Horváth et al. further examined Sim2Real domain randomization for robotic object detection [7]. Despite these advances, achieving consistent transfer across different industrial objects and real-world conditions remains unresolved.

Advances in visual architectures have also expanded the possibilities for industrial image recognition. He et al. introduced residual learning, enabling deeper convolutional networks to learn effective visual representations [6]. Tan and Le proposed EfficientNet, which systematically balances network depth, width, and image resolution to improve efficiency [15]. Dosovitskiy et al. introduced the Vision Transformer, demonstrating that Transformer-based architectures can perform image recognition using sequences of image patches [4]. Liu et al. later proposed a modern convolutional architecture designed to improve the capabilities of conventional convolutional networks [9]. These architectures provide different representation-learning characteristics, but their relative effectiveness under synthetic-to-real industrial conditions is not fully established. High performance within a synthetic domain may not necessarily correspond to strong generalization when the visual domain changes.

The existing literature therefore provides strong foundations for industrial recognition, synthetic data generation, domain randomization, and advanced visual architectures, but an important gap remains in their unified evaluation for sim-to-real industrial part classification. Many approaches concentrate on object detection, pose estimation, or individual transfer strategies rather than systematically examining classification performance across alternative deep learning architectures under a common synthetic-to-real setting. Furthermore, reliable industrial deployment requires predictions that can be interpreted rather than treated as unexplained outputs. The proposed system addresses this gap by bringing together synthetic-data-based visual learning, comparative architectural evaluation, real-world generalization, and explainable prediction within a unified framework for industrial part classification.

3. Materials And Methods

A. Proposed System

The proposed system presents a unified framework for sim-to-real industrial parts classification, designed to improve recognition when models trained on synthetic imagery are applied to real-world images. The framework encompasses synthetic data preparation, image preprocessing, deep learning model training, comparative evaluation, and real-world prediction. Multiple visual architectures are considered to identify an effective representation for cross-domain classification. The selected model is subsequently integrated with an explainable prediction mechanism to provide visual evidence supporting its decisions. A web-based interface enables users to submit industrial-part images and obtain classification results. By connecting synthetic-data learning with real-domain evaluation and interpretable prediction, the proposed system addresses the limitations of relying solely on synthetic-domain performance.

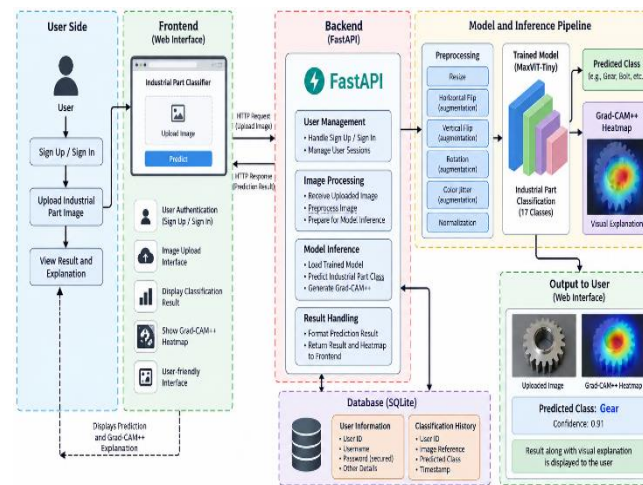


Fig. 1. System Architecture

Fig. 1 illustrates the overall architecture and information flow of the proposed system. Synthetic industrial images are first prepared and transformed into suitable inputs for model learning. The trained models subsequently process unseen images from the real domain for classification. The prediction stage produces the recognized class and supporting visual information through the explanation component. The final outputs are presented through the application interface, providing an integrated flow from data preparation and learning to real-world prediction and interpretation.

B) Dataset Collection

The Synthetic Industrial Parts dataset (SIP-17) was obtained from the publicly available Kaggle repository and is designed specifically for sim-to-real industrial parts classification. The dataset contains 17 industrial objects representing six typical industrial use cases, including isolated and assembled parts. It provides synthetic images with and without randomized backgrounds and post-processing, together with real images captured in industrial scenarios. The original dataset provides 1,200 synthetic training images and 300 validation images per object, while the real-world portion contains 566 test images. These characteristics make SIP-17 suitable for evaluating cross-domain generalization and the effectiveness of synthetic data for industrial visual recognition.

C) Image Pre-processing

The collected images are pre-processed to provide consistent inputs for deep learning models while preserving relevant visual information. Each image is resized to 224×224 pixels and converted into tensor representations. During training, data augmentation is applied through random horizontal and vertical flipping, rotation, and controlled color variation to increase visual diversity and improve model robustness. The images are subsequently normalized using standard ImageNet mean and standard deviation values. For validation and real-world testing, only resizing, tensor conversion, and normalization are applied, ensuring consistent and deterministic evaluation conditions.

D) Image Visualisation

Image visualisation is performed to examine the visual characteristics of the industrial-part dataset before model training. Representative samples are displayed from the available classes to observe differences in object shape, appearance, orientation, and surrounding context. Visual inspection also helps verify that the collected images correspond to the intended categories and provides an initial understanding of the variation present within the dataset. The visualised samples support qualitative assessment of the synthetic imagery and its suitability for subsequent classification and sim-to-real evaluation.

E) Algorithms

ResNet50: ResNet50 performs hierarchical image feature extraction through a deep residual architecture, enabling effective recognition of industrial parts. Its residual learning mechanism supports stable optimization and preserves informative features, contributing to high classification accuracy and robust visual representation.

ConvNeXt-Tiny: ConvNeXt-Tiny provides efficient visual feature learning through a modern convolutional architecture. It progressively transforms input images into discriminative representations for classification, offering a balance between computational efficiency, feature quality, and generalization across varying visual conditions.

EfficientNet-B0: EfficientNet-B0 performs image classification using a compact architecture designed to achieve strong recognition performance with relatively low computational requirements. Its balanced feature extraction process supports efficient learning of visual characteristics while maintaining competitive accuracy and generalization.

EfficientNet-B2: EfficientNet-B2 provides enhanced visual representation through a scaled convolutional architecture. It extracts increasingly detailed features from input images and uses these representations for classification, supporting accurate recognition while improving the ability to handle variations in object appearance and scale.

Swin Transformer Tiny: Swin Transformer Tiny captures both local and broader visual relationships through hierarchical attention-based feature learning. Its progressively structured representation enables effective recognition of object characteristics, supporting robust classification and improved generalization when visual patterns vary across image conditions.

ViT-B16: ViT-B16 processes images through a transformer-based representation that models relationships between image regions. It learns global visual characteristics for classification, providing an alternative to conventional convolutional architectures and supporting recognition through comprehensive contextual feature representation.

MobileNetV3: MobileNetV3 provides lightweight visual feature extraction for efficient image classification. Its compact architecture learns discriminative representations while reducing computational requirements, supporting practical recognition with a favorable balance between classification accuracy, inference efficiency, and model complexity.

CondConv EfficientNet: CondConv EfficientNet combines efficient convolutional feature extraction with adaptive representations that respond to different input characteristics. This approach enhances feature flexibility during classification, supporting improved recognition performance and robustness while retaining the efficiency associated with compact architectures.

MaxViT-Tiny: MaxViT-Tiny combines convolutional processing with attention-based feature learning to capture both localized details and broader spatial relationships. This integrated representation supports effective industrial-part recognition, contributing to strong classification accuracy and improved generalization across synthetic and real visual conditions.

Xception: Xception performs image classification through depthwise separable convolutional processing, enabling efficient extraction of meaningful visual features. Its architecture separates spatial and channel-wise feature learning, providing discriminative representations that support accurate recognition while maintaining computational efficiency.

F) Explainable Prediction

Explainable prediction is incorporated to improve the transparency and interpretability of the classification outcomes. The selected model generates a prediction for the input industrial-part image, while an associated visual explanation highlights the image regions that contribute most strongly to the predicted class. Grad-CAM++ is used to produce class-specific activation maps, which are transformed into heatmaps and overlaid on the original

image. This enables users to visually understand the basis of the classification decision rather than relying solely on the predicted label and confidence score. The resulting explanation supports easier interpretation of model behavior and provides greater confidence in automated industrial-part recognition.

G) Deployment

The developed classification framework is deployed as a web-based application to provide accessible and interactive industrial-part recognition. Users can access the application through an interface that supports account management and image submission. Uploaded images are processed and passed to the selected classification model for prediction. The application presents the predicted class along with its confidence information and provides a visual explanation of the decision using the generated activation map. This deployment enables users to interact with the classification system through a practical interface and facilitates the interpretation of automated predictions.

4. Experimental Results

A) Experimental Setup and Evaluation Metrics

The experimental setup evaluates the classification models under consistent training and testing conditions to assess their recognition capability and sim-to-real generalization. Images are standardized to a fixed input size and appropriate preprocessing and augmentation are applied during training. The models are trained using the synthetic image set and subsequently evaluated on both synthetic validation images and real-world images. Performance is measured using accuracy, precision, recall, and F1-score. Accuracy indicates overall classification correctness, while precision and recall assess prediction reliability and class recognition capability. The F1-score provides a balanced measure of precision and recall, enabling comprehensive comparison among the evaluated models.

Accuracy: Accuracy is the ability of a test to differentiate between healthy cases and patients. In each analysed case, the percentage of TP and TN should be computed to estimate an accuracy of a test. This can be expressed mathematically as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

Precision: Percentage of samples/incidents correctly classified as positive. Hence, the formula below can be employed to calculate the precision:

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (4)$$

Recall: In machine learning, recall is a metric that assesses a model's capacity to find all pertinent examples of a given class. It gives the information if a model is able to predict cases of a specific class very well and is defined as the proportion of positive cases predicted correctly out of all positive cases.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

F1-Score: F1 score is an evaluation metric used in machine learning, to assess the accuracy of a model. Combines the recall and precision scores for a model. The accuracy measure is the number of times the model was correct when predicting the entire data set.

$$F1 \text{ Score} = 2 * \frac{Recall \times Precision}{Recall + Precision} * 100 \quad (6)$$

B) Performance Evaluation

Table. 1. Performance Evaluation

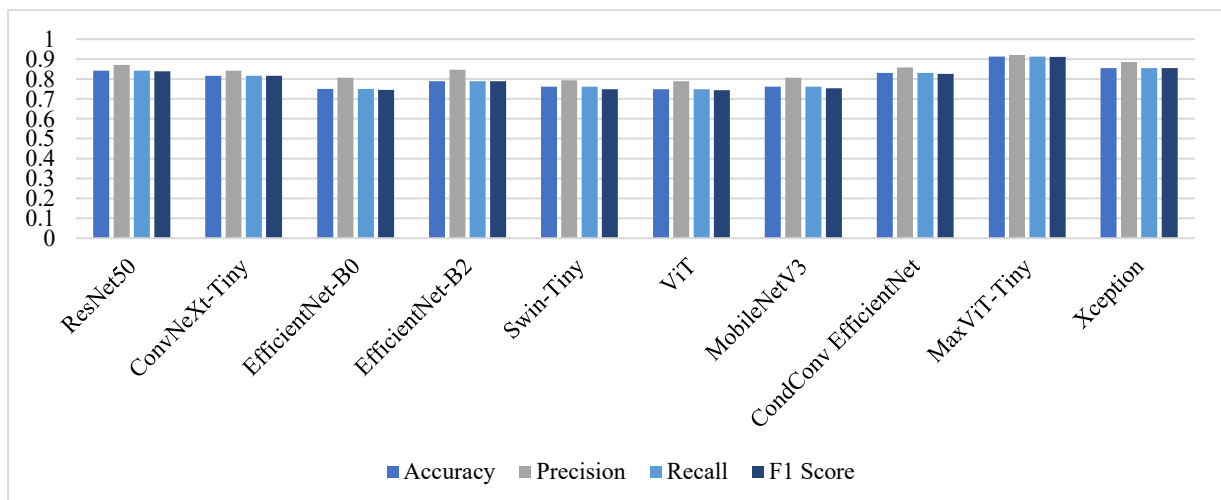
Model	Accuracy	Precision	Recall	F1 Score
ResNet50	0.842	0.871	0.842	0.839
ConvNeXt-Tiny	0.816	0.842	0.816	0.816
EfficientNet-B0	0.750	0.806	0.750	0.746
EfficientNet-B2	0.789	0.846	0.789	0.789
Swin-Tiny	0.761	0.794	0.761	0.748
ViT	0.749	0.788	0.749	0.743
MobileNetV3	0.761	0.806	0.761	0.754
CondConv EfficientNet	0.830	0.857	0.830	0.826
MaxViT-Tiny	0.912	0.920	0.912	0.911
Xception	0.855	0.885	0.855	0.855

The performance evaluation compares the classification models using four standard metrics on real-world test images. The results highlight differences in recognition capability and identify the model providing the most consistent overall classification performance.

The results show that MaxViT-Tiny achieves the highest performance across all four evaluation metrics, with 0.912 accuracy, 0.920 precision, 0.912 recall, and 0.911 F1-score. Xception and ResNet50 provide the next strongest results, while ViT and EfficientNet-B0 record comparatively lower performance. The consistently high scores of MaxViT-Tiny indicate stronger classification capability under real-world conditions and make it the best-performing model among the evaluated architectures.

C) Comparison Graph

The comparison graph provides a visual representation of the classification performance achieved by the evaluated models. It enables direct comparison across the selected evaluation metrics and clearly highlights the strongest and comparatively weaker-performing architectures.

**Fig. 3. Comparison Graph**

The graph demonstrates that MaxViT-Tiny consistently achieves the highest values across the evaluated metrics. Xception and ResNet50 show comparatively strong performance, whereas ViT and EfficientNet-B0 record lower values. The visual comparison confirms the superior real-world classification performance of MaxViT-Tiny.

D) Discussion

The evaluation results demonstrate clear differences in the ability of the models to generalize from synthetic imagery to real-world industrial-part images. MaxViT-Tiny achieved the highest overall performance, recording 0.912 accuracy and 0.911 F1-score, indicating consistent classification capability under real-world conditions. ResNet50 and Xception also produced comparatively strong results, while ViT and EfficientNet-B0 showed lower performance. The results suggest that architectural characteristics influence cross-domain recognition effectiveness. The performance variation across models further highlights the importance of evaluating classification approaches beyond synthetic validation results when assessing their practical suitability for sim-to-real industrial visual recognition.

E) Output Screenshots

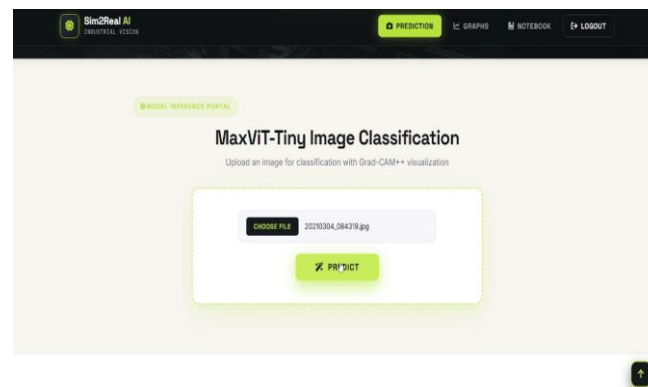


Fig.2 Input Interface

The image presents the MaxViT-Tiny classification interface, where users can upload an industrial-part image and initiate prediction with Grad-CAM++ visualization through the displayed prediction option.

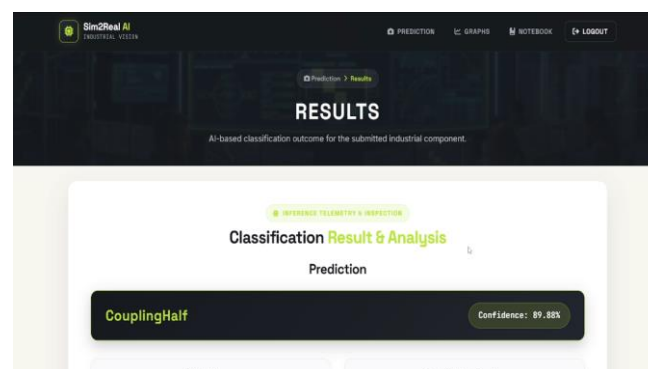


Fig.3 Classification Results

The image displays the classification output page, where the submitted industrial component is identified as CouplingHalf with a confidence score of 89.88%, indicating a reliable prediction result.

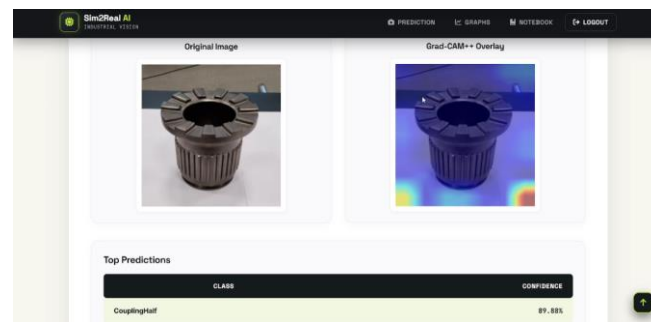


Fig.4 XAI Results

The image compares the original industrial component with its Grad-CAM++ overlay, highlighting influential visual regions while displaying the top prediction as CouplingHalf with an 89.88% confidence score.

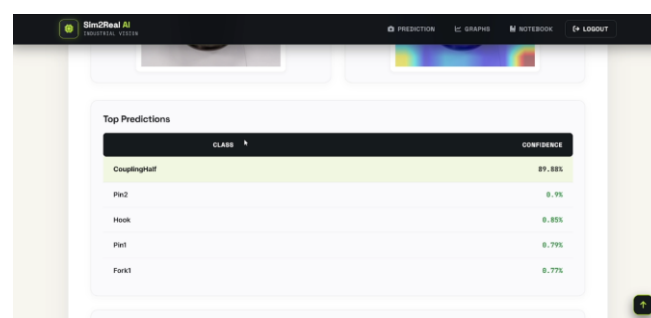


Fig.5 Top Predictions

The output displays the ranked top predictions, with CouplingHalf achieving the highest confidence of 89.88%, followed by Pin2, Hook, Pin1, and Fork1 with lower confidence scores.

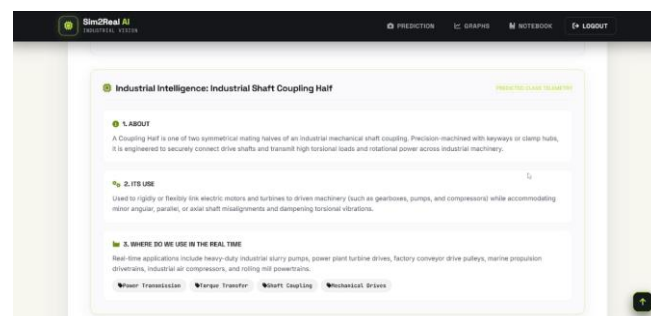


Fig.6 Detailed Information about Results

The output provides detailed information about the predicted CouplingHalf component, describing its industrial characteristics, intended usage, and real-time applications, supporting interpretation of the classification result and practical relevance effectively.

5. Conclusion

The developed system demonstrates that synthetic image data can provide an effective foundation for industrial part classification while reducing dependence on extensive real-world image collection. The comparative evaluation of ResNet50, ConvNeXt-Tiny, EfficientNet-B0, EfficientNet-B2, Swin Transformer Tiny, ViT-B/16, MobileNetV3, CondConv-EfficientNet, MaxViT-Tiny, and Xception established clear differences in their ability to transfer learned visual representations to real-world inputs. In particular, the evaluation confirmed that high performance on synthetic validation images does not necessarily guarantee equivalent performance under real-world conditions, emphasizing the importance of sim-to-real assessment. MaxViT-Tiny provided the most reliable classification capability among the evaluated architectures, achieving 91% accuracy on real-world test data. Its

selection, together with FastAPI-based deployment, resulted in a practical inference system capable of accepting user images and generating classification results. The integration of Grad-CAM++ further strengthens the system by making the classification decisions visually interpretable rather than relying solely on predicted labels. Overall, the developed solution demonstrates the feasibility of combining synthetic-data-based learning, robust visual classification, API-based deployment, and explainable AI for industrial inspection. It provides a practical basis for reducing manual component identification and supporting consistent, automated visual inspection in manufacturing environments.

Future development can focus on improving sim-to-real generalization by incorporating larger and more diverse real-world industrial image collections with variations in lighting, backgrounds, viewpoints, and component conditions. Domain adaptation and domain randomization techniques can further reduce the visual gap between synthetic and real-world data. Advanced transformer-based architectures and ensemble learning can be investigated to improve classification reliability. The system can also be extended to support real-time video-based inspection and integration with industrial cameras and robotic platforms. Further enhancement of Grad-CAM++ and other explainable AI techniques can provide more detailed visual interpretations. Deployment on edge devices can enable low-latency inspection in resource-constrained manufacturing environments.

References

- [1] J. Cohen, C. F. Crispim-Junior, C. Grange-Faivre, and L. Tougne, "CAD-based learning for egocentric object detection in industrial context," in Proc. 15th Int. Conf. Computer Vision Theory and Applications (VISAPP), vol. 5, 2020, pp. 644–651, doi: 10.5220/0008975506440651.
- [2] P. De Roovere, S. Moonen, N. Michiels, and F. Wyffels, "Dataset of industrial metal objects," arXiv preprint arXiv:2208.04052, 2022. The complete author list and dataset details were verified from the dataset source.
- [3] J. Dekhtiar, A. Durupt, M. Bricogne, B. Eynard, H. Rowson, and D. Kiritsis, "Deep learning for big data applications in CAD and PLM—research review, opportunities and case study," Computers in Industry, vol. 100, pp. 227–243, 2018, doi: 10.1016/j.compind.2018.04.005.
- [4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
- [5] L. Eversberg and J. Lambrecht, "Generating images with physics-based rendering for an industrial object detection task: Realism versus domain randomization," Sensors, vol. 21, no. 23, Art. no. 7901, 2021, doi: 10.3390/s21237901.
- [6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [7] D. Horváth, G. Erdős, Z. Istenes, T. Horváth, and S. Földi, "Object detection using Sim2Real domain randomization for robotic applications," IEEE Trans. Robotics, vol. 39, no. 2, pp. 1225–1243, 2023, doi: 10.1109/TRO.2022.3207619.
- [8] J. Krueger, J. Lehr, M. Schlueter, and N. Bischoff, "Deep learning for part identification based on inherent features," CIRP Annals, vol. 68, no. 1, pp. 9–12, 2019, doi: 10.1016/j.cirp.2019.04.095.
- [9] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986, doi: 10.1109/CVPR52688.2022.01167.
- [10] C. Mayershofer, D.-M. Holm, B. Molter, and J. Fottner, "LOCO: Logistics objects in context," in Proc. 19th IEEE Int. Conf. Machine Learning and Applications (ICMLA), 2020, pp. 612–617, doi: 10.1109/ICMLA51294.2020.00102.
- [11] X. Peng, B. Sun, K. Ali, and K. Saenko, "Learning deep object detectors from 3D models," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2015, pp. 1278–1286, doi: 10.1109/ICCV.2015.151.
- [12] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko, "VisDA: A synthetic-to-real benchmark for visual domain adaptation," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 2021–2026, doi: 10.1109/CVPRW.2018.00271.

- [13] A. Prakash, S. Bochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birchfield, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in Proc. IEEE Int. Conf. Robotics and Automation (ICRA), 2019, pp. 7249–7255, doi: 10.1109/ICRA.2019.8794443.
- [14] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in Proc. 14th European Conf. Computer Vision (ECCV), 2016, pp. 102–118.
- [15] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.
- [16] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in Proc. 16th European Conf. Computer Vision (ECCV), 2020, pp. 776–794.
- [17] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS), 2017, pp. 23–30, doi: 10.1109/IROS.2017.8202133.
- [18] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Bochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 969–977, doi: 10.1109/CVPRW.2018.00143.
- [19] M. Z. Wong, K. Kunii, M. Baylis, W. H. Ong, P. Kroupa, and S. Koller, "Synthetic dataset generation for object-to-model deep learning in industrial applications," PeerJ Computer Science, vol. 5, Art. no. e222, 2019, doi: 10.7717/peerj-cs.222.
- [20] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 2100–2110, doi: 10.1109/ICCV.2019.00219.