

Optimized Custom Object Classifier Using Augmented ResNet Vision Transformer Model for Defect Detection

Zia Ur Rehman, Wang Xing, Malak Abid Ali Khan, Yin Jing and Hongbin Ma*

National Key Lab of Autonomous Intelligent Unmanned Systems, School of Automation, Beijing Institute of Technology, Beijing 100081, China
mathmhb@bit.edu.cn

Abstract. The primary objective of this research is to enhance the computer vision tasks that are needed for specific object classification such as defected and effected often unable to deal with object variation imposed on by factors like different illumination, angles, and challenges. Traditional object classifiers experience a substantial challenge because of these variations, making it hard for computer vision models to correctly detect and classify objects in practical situations. In the proposed work we created an optimized strategy that includes an improved augmented vision model to get around these restrictions. Many advantages come with this upgraded Augmented ResNet vision transformer model, especially in terms of adapting to differences in unique object categorization. It is excellent at extracting features and expressing them, that leads to improved object detection and classification. To address the problem of dynamic variations in object appearance, our suggested model also incorporates a variation adaptive mechanism. Our model considers the dynamic nature of objects, including changes in their appearance and orientation, and adjusts its classification process accordingly. Through the utilization of this augmented Resnet vision transformer, our model effectively manages variations in object appearance, ultimately enhancing the accuracy of object classification. The combination of ResNet's convolutional features with Vision transformer self-attention mechanisms allows the mixed model to make more informed decisions based on a blend of local and global information, making it potentially more powerful and flexible for various computer vision tasks.

Keywords: Image classification, Image processing, Augmented model, ResNet Vision Transformer.

1 Introduction

Many industries, including the automotive, retail, and healthcare sectors, esteem image classification tasks. For tasks such autonomous driving, inventory management, and medical imaging analysis in these sectors, precise and effective object classification systems are crucial. Image classification systems are essential for identifying and classifying different objects on the road, such as pedestrians, vehicles, traffic signs, and obstacles [1]. These devices contribute to the occupants' and the autonomous vehicle's

safety. For activities like product identification and categorization in the retail sector, image classification is used. Inventory management requires accurate and effective object classification to guarantee that goods are accurately identified and arranged. In the healthcare sector, image classification systems are instrumental for disease diagnosis and medical image analysis [2]. The adoption of deep learning algorithms, particularly convolutional neural networks (CNNs), has brought about a revolutionary transformation in the realm of image classification. These networks have showcased exceptional performance in tasks such as image classification, object detection, and segmentation and their capacity to learn hierarchical representations from raw input data has proven highly effective in capturing and comprehending intricate visual patterns [3].

This development has led to significant improvement across industries where accurate and precise object classification is essential for decision-making and automation. For instance, deep learning-based image classification algorithms have been included in autonomous vehicles in the automobile industry to recognize and categorize diverse things on the road [4]. To analyse and understand real-time visuals, these systems use deep neural networks [5]. As a result, vehicles can make wise decisions and react to their environment. In the retail industry, deep learning algorithms have also been applied to create image classification systems that can accurately identify and categorize products. [6]. These systems find application in tasks like inventory management, where products must be precisely categorized and organized. Additionally, within the healthcare domain, deep learning algorithms have made significant strides in enhancing image classification systems. These systems contribute to precise disease diagnoses by analyzing medical images and detecting specific health conditions such as tumors and other abnormalities. Moreover, deep learning algorithms have extended their impact on histopathological image classification in breast cancer diagnosis [7].

The utilization of deep learning algorithms, notably CNNs, has demonstrated promising results in distinguishing between benign and malignant skin lesions and detecting Alzheimer's disease through MRI scans [8]. In recent years, both industrial and medical image analysis have witnessed substantial progress owing to the application of artificial intelligence algorithms, particularly deep learning. These algorithms have delivered state-of-the-art outcomes in various industrial and healthcare imaging tasks, including brain lesion segmentation, electrical and mechanical component sorting and placement, airway tree segmentation, diabetic retinopathy classification, and breast cancer metastasis detection [9]. Consequently, advancements in deep learning-based image classification have left a significant mark on diverse industries, including automotive, retail, and healthcare [10]. The accuracy and efficiency achieved by deep learning algorithms in image classification have revolutionized various industries.

It is hard to overestimate the significance of accurate image classification in today's world of rapid change. These developments have opened the way for innovation and progress in a range of industries in addition to streamlining several operations. Deep learning algorithms have revolutionized the categorization and detection of many diseases in the field of medical image processing [11]. A wide range of industries, including medical image processing, healthcare, and inventory management, accurate object classification is essential. Accurate image categorization plays a significant role in improving patient outcomes and assisting with diagnoses in the field of medical image

processing. Advancements in areas like healthcare and medical image processing have been made possible by deep learning algorithms, which have considerably improved the accuracy and efficacy of medical image classification task [12]. Additionally, the use of deep learning algorithms for the classification of medical images has shown success in improving the diagnosis and prognosis of disease. These algorithms have consistently produced state-of-the-art results in a range of medical imaging applications, including the segmentation of brain lesions, the identification of diabetic retinopathy, the segmentation of airway trees, and the categorization of cancer. Along with streamlining several procedures, these advancements have opened brand-new opportunities for innovation and growth throughout industries [13].

Accurate image classification acquires tremendous relevance, particularly in the medical field. In medicine, precise image classification has the potential to significantly improve disease diagnosis, direct therapy choices, and ultimately improve patient outcomes [14]. Additionally, it may assist in early disease detection, allowing for prompt action and a better outcome. Personalized treatment can also benefit from good image classification in medical image processing. Personalized medicine, that involves genetic profiles, medical histories, and scans, aims to customize medical treatments to individual patients based on their unique traits. This strategy may result in treatment plans that are more focused and exact, which would eventually improve patient's health and increase success rates. The development of computer-aided diagnosis systems can also benefit from accurate classification in medical image processing [15]. These systems employ image classification algorithms to assist healthcare professionals in making precise diagnoses. They can analyze medical images to detect anomalies, recognize patterns or features indicative of specific diseases, and provide supplementary insights to healthcare providers.[16]

In summary, precise image classification in the medical field harbors immense potential to enhance disease identification, prognosis, personalized medicine, and computer-aided diagnosis systems. While celebrating these advancements, it is imperative to acknowledge concerns regarding the reliability of deep learning systems in the context of medical diagnosis. In recent times, the introduction of transformers in image classification tasks, such as the Vision Transformer, has sparked interest in their potential for accurate classification in every field such as autonomous driving cars, natural language processing, and finance [17]. Since a long time ago, deep convolutional neural networks (CNNs) have been the preferred architecture for computer vision applications. In recent years, ResNets algorithm for image classification has been matched or even outperformed by transformer-based architectures like Vision Transformer (ViT). The usage of non-overlapping patches in the Transformer architecture, for example, makes one wonder if these networks are as resilient. The resilience of ViT models is thoroughly investigated in this research, and the results are contrasted with ResNet baselines. Both robustness to the model perturbations and robustness to the input perturbations are things we look into. We discover that ViT models, when pre-trained with enough data, are at least as resilient to a variety of perturbations as their ResNet counterparts [18].

Concerns about the reliability of deep medical diagnosis systems, particularly their vulnerability to adversarial attacks. For better disease identification in the medical field,

they suggest an augmented model that combines the advantages of Convolutional Neural Networks (CNNs) and Transformers. The suggested model makes use of Transformers' worldwide connectedness and CNNs' capacity to record local features. The authors employ effective convolution methods to reduce the large computational complexity of the Transformer's self-attention processes. Additionally, by learning smoother decision limits, they seek to strengthen the Transformer model's defense against hostile attacks. They present a method to do this, where they permute the feature mean and variance inside the high-level feature space to change the shape information of images [19].

2 Related Work

In recent years, there has been significant research and development in vision transformer architectures. Various approaches have been proposed to enhance the performance, efficiency, and architecture of vision transformers. These approaches include incorporating distillation strategies to improve training efficiency, introducing token-to-token modules and deep-narrow structures, developing multi-stage architectures with down-sampling, and exploring hierarchical network structures and multi-scale feature aggregation. Efforts have also been made to combine the advantages of 2D CNNs with transformers by incorporating convolutional layers [20]. One of the primary challenges in training vision transformers is their computational and memory requirements. Researchers have devised distillation strategies to address these challenges. Distillation involves transferring knowledge from a large, pre-trained model to a smaller one. This strategy speeds up training while reducing the computational load. To increase the effectiveness of training, strategies like knowledge distillation, attention distillation, and feature distillation have been investigated. To match the inference efficiency of traditional CNNs, studies on developing efficient and lightweight topologies for networks, such as MobileNets, EfficientNets, and ShuffleNets, have also been done [21].

Research in computer vision has mostly focused on effective network topologies to improve the performance and computational efficiency of vision tasks. In recent years, convolutional neural networks have undergone substantial research to boost their efficiency and performance in vision-related applications. The field of computer vision has recently placed a lot of emphasis on creating effective network topologies [22]. The need to improve the performance and computational effectiveness of visual tasks is what motivates this focused research. The investigation and improvement of efficient and light-weight network architectures have taken center stage among the many strategies investigated. MobileNets, EfficientNets, and ShuffleNets are a few notable contributions in this field, each of which offers distinctive insights and breakthroughs. One of the first innovations in the search for effective network topologies, MobileNets, has been instrumental in changing the face of computer vision. These networks were created especially to address the resource limitations of embedded and mobile devices, where conventional convolutional neural networks (CNNs) frequently fail because to high computing needs [23].

Depth-wise separable convolutions are used by MobileNets to achieve efficiency. These networks use a two-step method that combines pointwise and depth-wise

convolutions rather than performing a standard convolutional operation on all input channels. This lowers the number of calculations and parameters while maintaining the capacity to learn sophisticated features. The end result is a thin network architecture that still performs various vision tasks very well [24].

Another significant step has been taken in the search for effective network topologies with EfficientNets. They present a novel strategy that methodically increases the model's depth, width, and resolution while preserving a constant trade-off between model size and performance. This scalability enables researchers to fine-tune models to satisfy particular computational constraints without substantially compromising accuracy. Each of the variants in the EfficientNet family, including B0 through B7, offers a different trade-off between computational cost and performance. This adaptability has made EfficientNets a well-liked option for a variety of applications, from object detection to image classification. Although the development of MobileNets, EfficientNets, and ShuffleNets has made significant strides in the field of efficient network design, it is crucial to emphasize that convolutional neural network (CNN) research is still in its infancy. The efficiency and performance of CNNs for various vision applications are constantly being improved by researchers. [25].

The creation of novel convolutional operations and architectures that enhance the receptive field, feature extraction, and gradient flow within networks is one prominent trend. The incorporation of transformer-inspired attention mechanisms into CNNs has also improved the modeling of long-range dependencies in visual data. Computer vision has greatly advanced thanks to the concerted efforts to develop effective network topologies, which are exemplified by innovations like MobileNets, EfficientNets, and ShuffleNets. These architectures have significantly increased the inference efficiency of vision tasks and made it possible to deploy vision models on a variety of devices, including mobile phones, edge devices, and other devices.

[26] Additionally, the constant quest to improve the efficiency and performance of conventional CNNs continues to be a driving force in computer vision research. This ongoing innovation makes sure that the field keeps pushing the limits of what is possible in visual understanding and recognition, opening the door for fascinating developments in the years to come. In essence, the pursuit of effectiveness and performance in computer vision continues to be a dynamic and ever-evolving process with plenty of room for advancements. [27].

One of these accomplishments is the development of efficient network topologies like MobileNets, EfficientNets, and ShuffleNets, which aim to match the inference efficiency of traditional CNNs while using less CPU power. Recent advances in vision transformer topologies have greatly enhanced performance, efficiency, and overall architecture. Recent advances have been made in vision transformer architectures' performance, efficacy, and design. Recent advancements in vision transformer designs have revolutionized the field of computer vision, leading to significant improvements in architecture, performance, and efficiency [28].

3 Methodology

The Vision Transformer is a powerful model that has shown remarkable performance in image classification tasks. To utilize the Vision Transformer for a classification task, several steps need to be taken.

3.1 Data Processing

The first step in using a Vision Transformer for classification is to prepare the data. This involves collecting and preprocessing a dataset of images that are labeled with their corresponding classes. This dataset should be diverse and representative of the classes that the model will be trained on. Additionally, it is essential to ensure that the dataset is properly split into training, validation, and test sets to evaluate the performance of the model accurately [29]. In the proposed model, there are 20 classes mentioned by name as WP-1, WP-2,, WP-20 of different objects for the classification task as shown in Fig. 1.



Fig. 1. 20 different classes for objects Classification

This small dataset contains different mechanical components and chess pieces of the Chinese chess game, we take 200 images of each class with different angles for variation adaptation and labeling the dataset shown in Fig. 2.

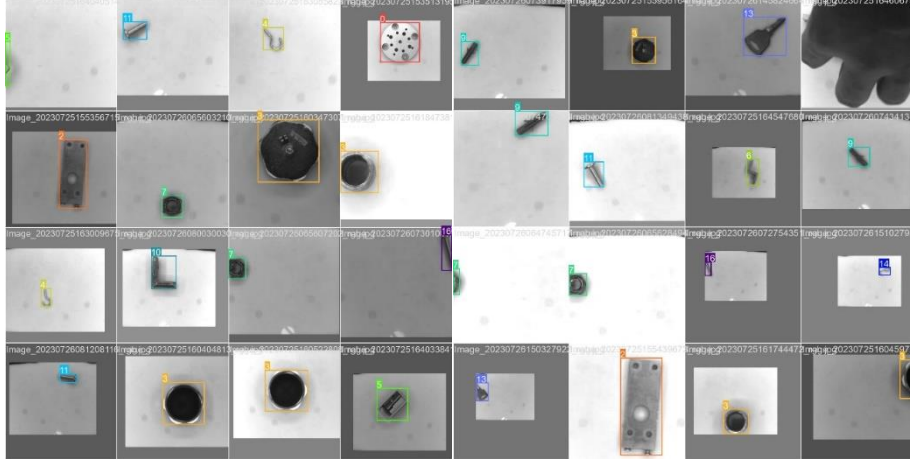


Fig. 2. Labeling the dataset

3.2 Model Architecture

To define the architecture of the Augmented ResNet Vision Transformer model. ResNet networks like ResNet-50, ResNet-101, and ResNet-152 are trained for object recognition and can be used as generic feature extractors to derive features for a standard classifier. These ResNet networks have been developed to achieve human-level performance tasks by computer vision, such as image recognition, which has significantly improved by the powerful representational ability of ResNet networks.

In this work, we use the Res-Net architecture as the feature extractor. Choosing it as the backbone architecture of our feature extraction process. ResNet is a powerful CNN architecture that can extract important visual features. We input these features to the transformer network for further processing and analysis after extracting. First, we trained the architecture then convert them into a grid of smaller size which depend on the size of the patch. In the proposed work the feature maps are pooled down to a 7X7 grid of patches which are then linearly embedded to a hybrid ResNet vision transformer model. This involves specifying the number of layers, the dimensionality of the model's embeddings, and the size of the patches that will be extracted from the input images. The Vision Transformer architecture is a recent development in computer vision that adapts the original Transformer architecture, which was initially designed for natural language processing tasks, to vision classification problems [30]. Unlike traditional convolutional neural networks, which rely heavily on convolutions and pooling operations, the Vision Transformer utilizes self-attention mechanisms to capture long-range dependencies and relationships between different elements in an image [31]. This architecture consists of several building blocks that work together to achieve high-performance image classification. Let $S = \{w_i, y_i\}_{i=1}^n$ denotes a set of work piece model images, w_i is the image of the workpiece model and y_i is the corresponding output label for each class. The first building block of the augmented Res-Net-Vision Transformer architecture is the feature extraction and patch embedding, which converts an

input image into a sequence of small patches in the proposed model we split the image into 8, 12, and 16 patches and analyzed the result and found that 12 patches show better result for our task. These patches serve as the input tokens for the subsequent layers of the Res-Net Vision Transformer as shown in Fig. 3. Each patch is then linearly embedded into a higher-dimensional feature space, allowing the model to capture more complex and abstract visual information [32]. The next building block is the positional encoding, which provides information about the spatial position of each patch. This helps the model understand the spatial relationships between patches and maintain an understanding of the overall image structure. The positional encoding is added to the patch embeddings before they are passed through the Transformer layers. The core components of the Vision Transformer design are the Transformer layers. They are made up of numerous feed-forward and self-attentional sub-layers. Each patch embedding interacts with every other patch through a self-attention mechanism in the self-attention sub-layers [33]. This enables the model to concentrate on various areas of the image and identify important spatial connections. The self-attention mechanism computes a weighted sum of the values generated by hidden layers while considering the relative relevance of various patches. The original patch embeddings are then combined with these weighted sums to give the model information from different parts of the image. The Transformer layers' feed-forward sub-layers are responsible for performing non-linear transformations on the patch embeddings. These modifications enable the model to capture complex and advanced visual aspects [34].

The Vision Transformer splits the input image into patches and applies a linear projection to each patch, resulting in a fixed-sized vector. Here in this work instead of image pacification, we introduce a novel idea that first extracts the important features through ResNet pacifies these filtered featured images to a vision transformer for pacification, and then feeds these featured patches to a linear embedding layer for position embedding, as positional encoding, representing the spatial location of each patch, is then concatenated with the patch embeddings. The Vision Transformer architecture for image classification tasks consists of several key steps that allow the model to effectively capture long-range dependencies and global context in images [35].

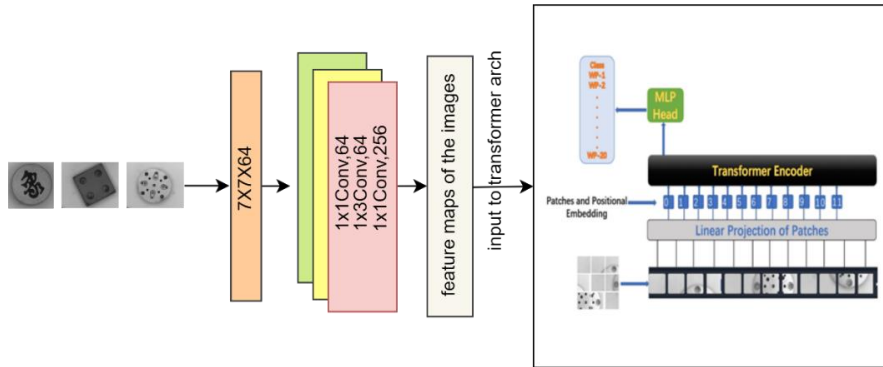


Fig. 3. Schematic diagram of Augmented ResNet vision transformer model architecture

3.3 Training the Model

Once the data and model architecture are prepared, the Res-Net-Vision Transformer can be trained on the labeled dataset. Training the Vision Transformer involves feeding the image patches from the dataset into the model and updating its parameters using an optimization algorithm “Adam”. Algorithm 1 shows the key steps for training the Vision Transformer. In the propose work we train our model for different sets of classes that is, for 2, 4, 8, 16, and 20 classes, and observe the efficiency and accuracy of the model as shown in Fig. 4.

Algorithm 1

Input: Training Data: $S = \{w_i, y_i\}_{i=1}^n$

Output: feature map

Input: feature images

Output: Predict the class of each test image

1 Set batch size to 50 optimize Adam with the learning rates (0.0002,0.0003,0.0004) Number of iterations 20,30,40 and set image dimension to 250X200.

2 Number of mini-batches; $N_b = N/\text{Batch Size}$

3 For iteration = 1: N

- For Batch = 1: N_b
- Select a batch from the training data set.
- Choose another batch from the augmented data set.

4 Train the model on real and augmented images

- Backpropagate the loss.
- Perform backpropagation through the network to compute gradients concerning the model's parameters.
- Update the model parameters.

Predict the class

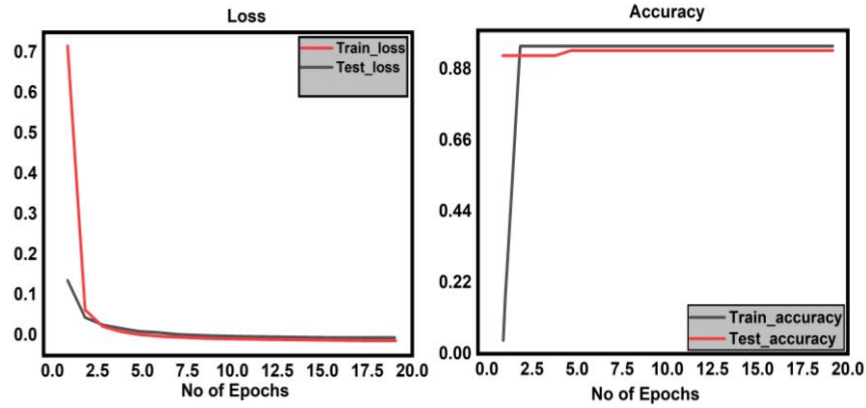


Fig. 4. Training Loss and Accuracy for 20 Epochs with Learning Rate=4

3.4 Evaluation

After training, the performance of the Vision Transformer needs to be evaluated. This can be done by using the test set that was previously set aside. The Vision Transformer model generates predictions for the test images and compares them to the ground truth labels. The custom data set of 20 different classes is used for the evaluation of the model which shows efficient results as shown in Fig. 5.

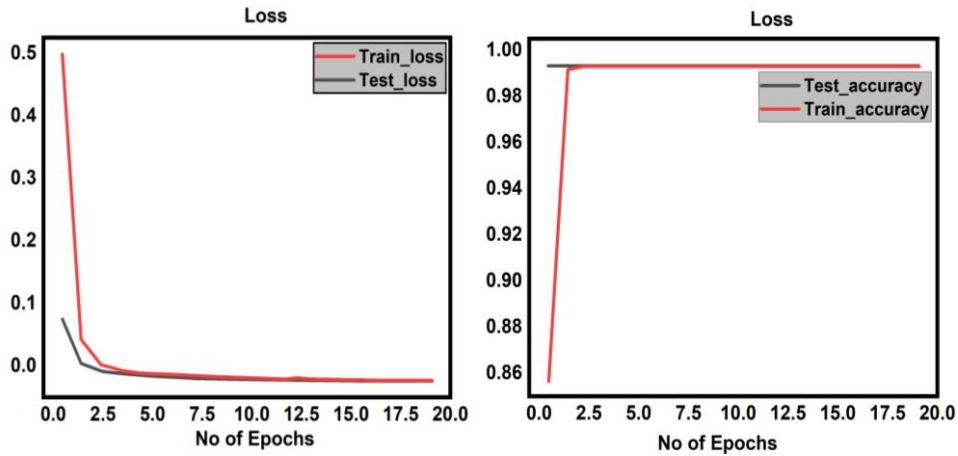


Fig. 5. Training Loss and Accuracy for 20 Epochs with Learning Rate=3

3.5 Fine Tuning

In some cases, it may be necessary to fine-tune the Vision Transformer model. This can be done by unfreezing certain layers of the model and training it on a smaller, task-specific dataset. This process allows the model to adapt to the specific requirements of the classification task and further improve its performance beside this we also apply different learning rates to adjust the training loss and overfitting as the learning rate is proportional to overshoot and divergence, if the learning rate is too high, the network may overshoot from the optimal solution and diverge and the network may converge too slowly or get stuck in a local minimum if the learning rate is too low. To avoid overfitting or underfitting we conducted several experiments with different sets of classes and different learning rates and found that 0.0003 is the optimal learning rate for our task as shown in Fig. 6. The overall performance of the model is given in Fig. 7.

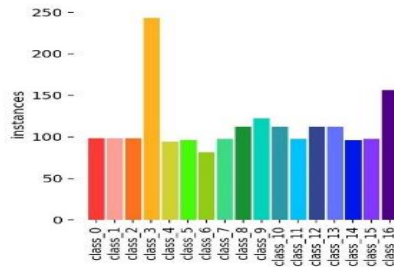


Fig. 6. Variation of instances for classes

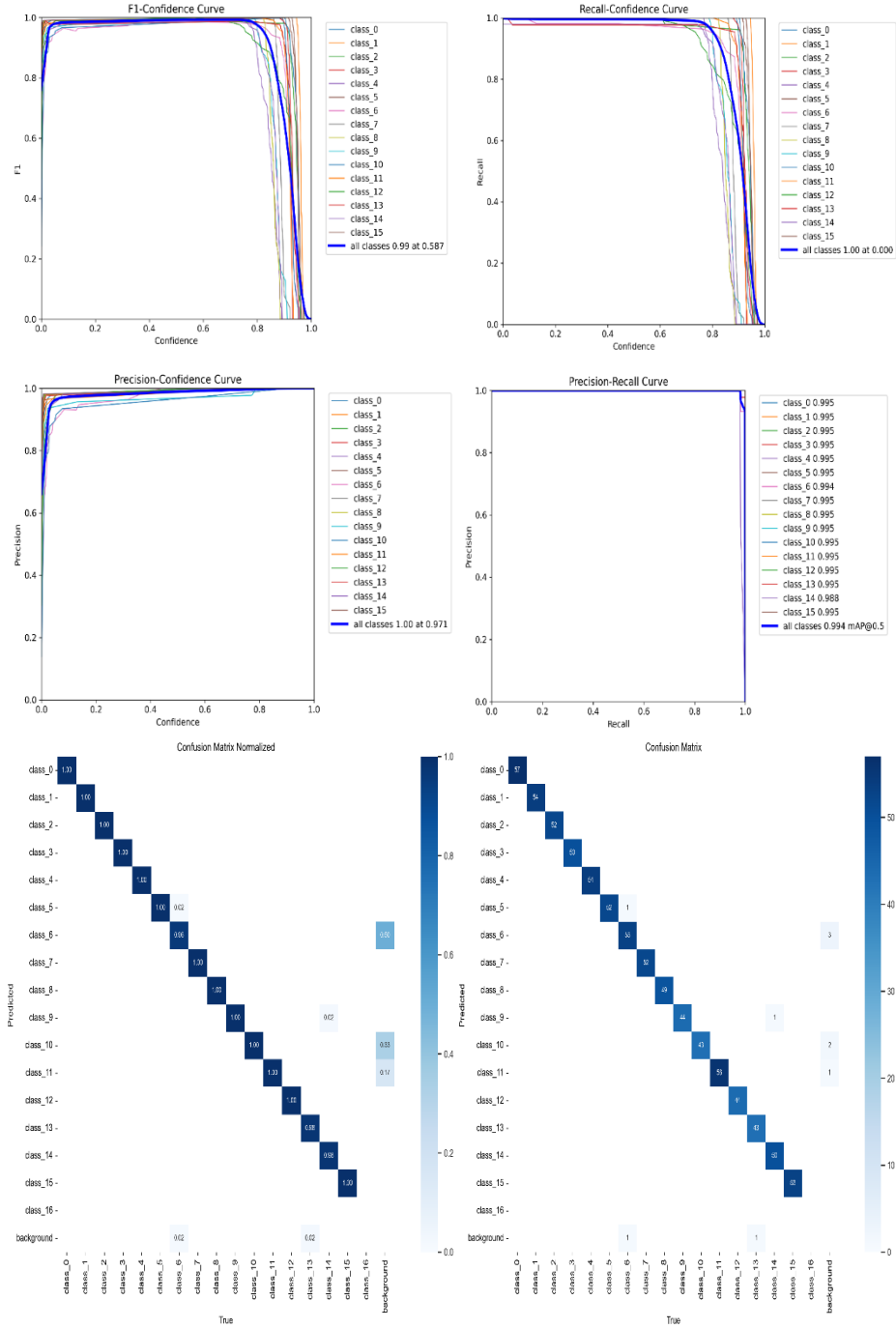


Fig. 7. Performance of the model














3.6 Deployment








Finally, the trained and fine-tuned Augmented ResNet-Vision Transformer model is deployed for practical use, which shows good results as compared to the other models. The accuracy for different classes at the fixed image and patch size is shown in Table 1, while Table 2 illustrates the true and predicted labels for each class.

Table 1. Accuracy for different classes

No of classes	Image size	Patch size	accuracy
2	250 X 200	32	100%
4	250 X 200	32	99%
8	250 X 200	32	99.5%
16	250 X 200	32	98%
20	250 X 200	32	98%

Table 2. Label Prediction

	Class	Actual Label	Predicted label
	Round Work Piece	Wp-1	Wp-1
	Square Work Piece	Wp-2	Wp-2
	Rectangle Work Piece	Wp-3	Wp-3
	Hex Bolt	Wp-4	Wp-4
	Threaded Huck	Wp-5	Wp-5
	Sharpener	Wp-6	Wp-6
	Double L Bend	Wp-7	Wp-7
	Nut	Wp-8	Wp-8
	Square Washer	Wp-9	Wp-9
	Screw Bolt (big)	Wp-10	Wp-10
	L Bend	Wp-11	Wp-11
	Screw Bolt (small)	Wp-12	Wp-12
	Drill Bit	Wp-13	Wp-13

	Key	Wp-14	Wp-14
	Stapler Pin	Wp-15	Wp-15
	Chinese Character Horse	Wp-16	Wp-16
	Chinese Character Elephant	Wp-17	Wp-17
	Chinese character Canon	Wp-18	Wp-18
	Chinese character Vehicle	Wp-19	Wp-19
	Chinese character Soldier	Wp-20	Wp-20

4 Conclusion

This research introduces an innovative approach by augmenting the pre-processing layers of ResNet into the patch embedding layer for self-supervised training which give and efficient results. We assess its impact on industrial part/component image classification using the augmented model. Our findings, validated on a custom dataset of 20 industrial object classes, demonstrate enhanced performance. This approach holds promise for accurate industrial object identification and underscores the significance of adaptive models in vision transformers. The combination of ResNet's convolutional features with ViT's self-attention mechanisms allows the mixed model to make more informed decisions based on a blend of local and global information, making it potentially more powerful and flexible for various computer vision tasks.

Acknowledgments. This work was funded by the National Key Research and Development Plan of China (2018AAA0101000) and the National Natural Science Foundation of China (62076028).

References

1. Soori, M., Arezoo, B., & Dastres, R. (2023). Artificial intelligence, machine learning, and deep learning in advanced robotics: A review. *Cognitive Robotics*.
2. Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsafaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., & Pattichis, C. S. (2020). AI in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, 24(7), 1837-1857.
3. Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, 9, 611-629.

4. Gupta, A., Anpalagan, A., Guan, L., & Khwaja, A. S. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10, 100057.
5. Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O., & Ahmed, A. A. (2020). Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, 32, 101827.
6. Hyun, Y., & Kim, D. (2022, June 21). Development of deep-learning-based single-molecule localization image analysis. *Journal of Imaging*.
7. Debelee, T. G., Kebede, S. R., Schwenker, F., & Shewarega, Z. M. (2020). Deep learning in selected cancers' image analysis—A survey. *Journal of Imaging*, 6(11), 121.
8. Sarvamangala, D. R., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: A survey. *Evolutionary Intelligence*, 15(1), 1-22.
9. Najjar, R. (2023). Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17), 2760.
10. Valente, J., António, J., Mora, C., & Jardim, S. (2023). Developments in image processing using deep learning and reinforcement learning. *Journal of Imaging*, 9(10), 207.
11. Cai, L., Gao, J., & Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. *Annals of Translational Medicine*, 8(11).
12. Amin, J., Sharif, M., & Yasmin, M. (2016). A review on recent developments for detection of diabetic retinopathy. *Scientifica*, 2016.
13. Wen, Z., & Huang, H. (2022). The potential for artificial intelligence in healthcare. *Journal of Commercial Biotechnology*, 27(4).
14. Mathur, S., & Sutton, J. (2017). Personalized medicine could transform healthcare (Review). *Biomedical Reports*, 7(1), 3-5.
15. Javaid, M., Haleem, A., Singh, R. P., Suman, R., & Rab, S. (2022). Significance of machine learning in healthcare: Features, pillars, and applications. *International Journal of Intelligent Networks*, 3, 58-73.
16. Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9), 5521.
17. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., & Veit, A. (2021). Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10231-10241).
18. Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157, 106791.
19. Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer-based variants. *Artificial Intelligence Review*, 1-54.
20. Gao, M. (2023). A survey on recent teacher-student learning studies. *arXiv preprint arXiv:2304.04615*.
21. Taye, M. M. (2023). Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. *Computation*, 11(3), 52.
22. Dong, F., Qiu, C. W., Qiu, J., Hua, K., Su, W., Wu, J., & Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research.
23. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

24. Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning* (pp. 6105-6114). PMLR.
25. Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8(3), 331-368.
26. Campos Zabala, F. J. (2023). Neural networks, deep learning, foundational models. In *Grow Your Business with AI: A First Principles Approach for Scaling Artificial Intelligence in the Enterprise* (pp. 245-275). Berkeley, CA: Apress.
27. Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22), 4712.
28. Arshed, M. A., Mumtaz, S., Ibrahim, M., Ahmed, S., Tahir, M., & Shafi, M. (2023). Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models. *Information*, 14(7), 415.
29. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 54(10s), 1-41.
30. Khan, A., Rauf, Z., Sohail, A., Rehman, A., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and its CNN-transformer based variants. *arXiv preprint arXiv:2305.09880*.
31. Chen, C. F. R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 357-366).
32. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2010). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
33. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., & Feng, J. (2021). Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*.
34. Al-Hammuri, K., Gebali, F., Kanan, A., & Chelvan, I. T. (2023). Vision transformer architecture and applications in digital health: A tutorial and survey. *Visual Computing for Industry, Biomedicine, and Art*, 6(1), 1-28.