IJDASEA
International Journal of Data Science Engineering, and Analytics

e-ISSN : 2807-1689
p-ISSN : 2798-9208

# Wayang's Images Recognition using Vision Transformer

**Andreas Nugroho Sihananto [1,*], Muhammad Muharrom Al Haromainy [2] Zaky Ahmad Fauzi [3], Reno Alfa Reza [4], Gredy Christian Hendrawan Putra [5] and Theressa Marry Christianty [6]**

Vol. 04, No 02, Page 15-27.
Received: 18 Nov 2024
Accepted: 19 Nov 2024
Published: 2 Dec 2024

[1] Affiliation 1; andreas.nugroho90@gmail.com
[2] Affiliation 2; muhammad.muharrom.if@upnjatim.ac.id
[3] Affiliation 3; 22081010282@student.upnjatim.ac.id
[4] Affiliation 4; 23081010091@student.upnjatim.ac.id
[5] Affiliation 5; 22081010195@student.upnjatim.ac.id
[6] Affiliation 6; 23081010131@student.upnjatim.ac.id
* Correspondence: andreas.nugroho90@gmail.com

*Due to its complex nature and outdated perception, Wayang is a traditional Indonesian art form influenced by Hindu-Buddhism. However, it is difficult for the younger generation to recognize the various types of Wayang. In an effort to preserve Wayang culture, this study evaluates the performance of four deep learning models in recognizing types of Wayang namely, Vision Transformer (ViT), ResNet34, YOLOv5-cls, and YOLOv8-cls. These models were trained and assessed using a dataset of 232 images representing six Wayang types and using metrics ssuch as accuracy, recall, precision, and F1 score. ViT demonstrated efficiency and adaptability despite high computational requirements, achieving the best accuracy (91.3%). Meanwhile, YOLOv5-cls and YOLOv8-cls offered a good balance between accuracy and efficiency. This study suggest that deep learning models can play an essential role in preserving Wayang culture by improving its recognition and accessibility, thus helping younger generations appreciate this traditional art form.*

**Keywords:** Wayang, deep learning, vision transformer, residual network-34, yolov5-cls, yolov8-cls, image classification

## 1. Introduction

Wayang is one of Indonesia's traditional art forms that has been rooted since the time of our ancestors. According to the Indonesian Dictionary, Wayang refers to puppets made of carved leather or wood, used to portray characters in traditional drama performances in regions such as Bali, Java, and Sunda. The origins of Wayang can be traced back to around the 5th century BC, during which Indonesians practiced animism and dynamism. Hindu-Buddhist influences from the 6th

10.33005/ijdasea.v4i2.24

century AD enriched Wayang stories with the Mahabharata and Ramayana epics, leading to its development into an inseparable part of Indonesia's cultural heritage [1]. Various types of Wayang have emerged in Indonesia, including Wayang Kulit, Wayang Golek, Wayang Gedog, Wayang Beber, and Wayang Suket.

As a cultural heritage, Wayang risks extinction if there is insufficient appreciation from the community and government towards Wayang artists [2]. Such support is crucial for enhancing the enthusiasm and motivation of Wayang artists to continue their work and preserve this art form. The dalang association also plays an important role in engaging the younger generation and exploring ways to preserve Wayang [3]. Other efforts to preserve Wayang include organizing performances during significant occasions and holding competitions related to Wayang art.

Research indicates several factors that contribute to the younger generation's difficulty in recognizing different types of Wayang, such as complex characterization, outdated and boring perceptions, and a lack of knowledge about various Wayang types [3]. Furthermore, online search engines often yield irrelevant results when identifying Wayang, especially regarding types from specific regions. Conventional image search technologies have proven ineffective in recognizing and identifying different types of Wayang, creating challenges for users seeking accurate information. Therefore, a modern approach is needed to address the issue of automatically identifying types of Wayang.

One potential solution is the application of deep learning detection technologies. This research aims to compare the effectiveness of the Vision Transformer model against Residual Network and YOLO variants in identifying Wayang types effectively and efficiently. Ultimately, this study seeks to encourage the younger generation to learn about and inherit the Wayang culture.

## 2. Related Works

Several research studies are discussing the meeting point of the deep learning technologies in the scope of classification of Wayang, and each one contributes different insights and methodologies to this rising field. One of the prominent works by [26] has used the YOLOv5 algorithm to detect the Balinese shadow wayang character only in the "Wayang Peteng" performance. The study achieved great results, where the YOLOv5n model, trained for 200 epochs, reached perfect precision and recall values, as well as a mean Average Precision (mAP) at a threshold of 0.5 of 0.995, reaching 128.20 frames per second. Such really high values of those metrics definitely outline the effectiveness of YOLOv5 in a specific context, but one limitation exists in this research: it is single-type performance-oriented. Realizing this gap, our research is on the way to extend the dataset to Wayang Gedog, Wayang Golek, Wayang Kulit, and others. By

doing so, we hope to get a more holistic analysis of Wayang classification. We also compare the performances of different object detection methods, namely, ViT, ResNet, YOLOv5 Classification, and YOLOv8 Classification, to determine which method performs better and is more efficient in detecting and classifying these various types of Wayang.

In another study, [27] did a comparative study between Convolutional Neural Networks (CNNs) and Residual Networks (ResNet) in classifying the severity level of diabetes. The result showed that ResNet performed better than CNN, with an accuracy of 81.23% as opposed to CNN's 68.49%. This proves that ResNet is much stronger in handling complicated classification tasks, therefore substantiating our choice in using ResNet as a base object detection method in our research. We believe that harnessing the proven accuracy of ResNet can greatly improve our initiative of classifying these different types of Wayang, which often show complex features that are vital in thoroughly distinguishing them.

Furthermore, another study of great importance by [28] compared the strengths of the Vision Transformer (ViT) to CNN in properties such as the DeiT-Tiny ViT model versus ResNet18 for tumor detection and tissue type identification. The results is ViT slightly outperformed ResNet18 in tumor detection but was outperformed by ResNet18 in network-type identification. These results demonstrate that the ViT architecture, with its potential for dealing with difficult visual tasks, outperforms more traditional CNN architectures. However, its really unique architecture gives it an advantage over CNNs in scenarios where complex image representations are needed.

This finding strongly supports our decision to incorporate ViT in our framework for classifying Wayang types, since its advanced self-attention mechanisms are capable of relationships between different image components, this being a very important aspect in recognizing the different styles of Wayang. Overall, the related work shows a fertile landscape of exploration that provides a strong foothold for our study. It addresses the drawbacks from previous research by using multiple advanced detection methodologies to fill up the gaps in the literature and help contribute to better knowledge of Wayang art forms through innovative computational means.

**3. Experiment and Analysis**

This research on Wayang image recognition involved training several deep learning models, including ResNet34, YOLOv5-cls, YOLOv8-cls, and Vision Transformer (ViT), on a dataset of 232 six types of Wayang. The dataset was preprocessed by resizing, augmentation, and splitting into training, validation, and test sets. Models were evaluated based on the main metrics of accuracy,

recall, precision, and F1-score, as well as efficiency and speed, since the application is for real-time tasks. ResNet34 was a CNN with residual blocks, good at capturing features but computationally intensive. On the other hand, YOLOv5-cls and YOLOv8-cls brought fast and efficient classification, suited for real-time needs; however, they were not good at dealing with complex image details. Similarly, ViT implemented a Transformer architecture for capturing detailed relationships among image patches, therefore achieving a high, generalized accuracy.

*3.1. Dataset*

The dataset used in this study contains a collection of images of 6 types of puppets, namely Wayang Gedog, Wayang Golek, Wayang Krucil, Wayang Kulit, Wayang Suluh, and Wayang Beber. The dataset used in this research is a dataset obtained directly through an online-based dataset provider source, namely Kaggle and moved to https://app.roboflow.com/. Quick access to get the dataset can be accessed through the following page https://universe.roboflow.com/Wayang-cumb2/Wayang-classification. The image set consists of 34 Wayang Gedog images, 40 Wayang Golek images, 41 Wayang Krucil images, 45 Wayang Kulit images, 33 Wayang Suluh images, and 39 Wayang Beber images. There are 232 image datasets with 224x224 pixels stored in .jpg format. This research uses a single label method in classification on each dataset.


**Figure 1** Types of Wayang.

*3.2. Methodology*

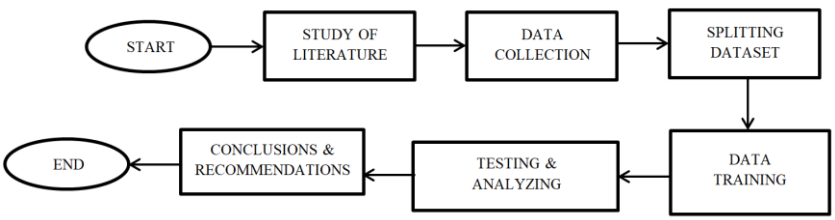


**Figure 2** Flowchart of Research Stages.

The data processing process starts with changing the image resolution to 224x224 pixels to ensure the consistency of the model input. Next, data augmentation techniques such as scaling, flipping, and rotation are applied to reduce overfitting and increase dataset variety. Pixel normalization is also performed to accelerate convergence during training. The dataset was then divided into subsets for training (70%), validation (20%), and testing (10%) to ensure objective evaluation. The training model involves several deep learning architectures, including ResNet34, YOLOv5-cls, YOLOv8-cls, and Vision

Transformer (ViT), each of which has advantages in image classification. During training, various parameters are set, including the number of epochs and learning rate, using optimization algorithms such as SGD and ADAM. Model evaluation is performed using metrics such as accuracy, recall, precision, and F1-Score, as well as cross-validation techniques to ensure the model can adapt well to various conditions. After training and evaluation, the model is tested using different data from the training data to ensure the accuracy and ability of the model to recognize Wayang types that have never been seen before.

**Table 1** Experimental Design.

| Architecture Model | Training Data | Testing Data | Validation Data | Learning Rate | Epoch |
|---|---|---|---|---|---|
| YOLOv5-cls | 70% | 10% | 20% | 0.001 | 50 |
| YOLOv8-cls | 70% | 10% | 20% | 0.001 | 50 |
| ResNet34 | 70% | 10% | 20% | 0.002 | 50 |
| Vision Transformer | 70% | 10% | 20% | 0.0002 | 50 |

*3.3. ResNet34 model applied to Wayang classification*

These are the testing results and confusion metrics rate for the ResNet34 model applied to Wayang classification.

**Table 2** ResNet34 model applied to Wayang classification.

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Beber | 1.0 | 1.0 | 1.0000 |
| Gedog | 1.0 | 1.0 | 1.0000 |
| Golek | 1.0 | 0.75 | 0.8571 |
| Krucil | 0.5 | 0.5 | 0.5000 |
| Kulit | 0.66 | 0.8 | 0.7272 |
| Suluh | 1.0 | 1.0 | 1.0000 |
| Overall | | | 0.8260 |

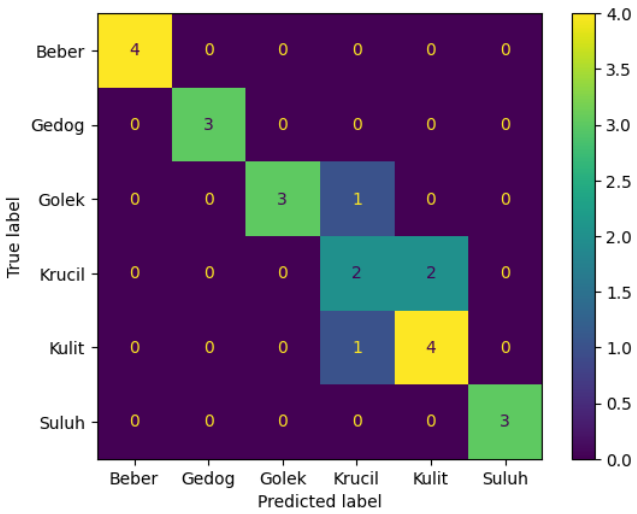


**Figure 3** Confusion Metrics of ResNet34.

The ResNet34 model performed well overall, excelling in classifying Wayang Gedog, Wayang Suluh, and Wayang Beber with perfect accuracy. Wayang Kulit and Wayang Golek were classified reasonably well with few errors, but Wayang Krucil showed less satisfactory performance, with moderate recall and precision. The confusion metrics indicates that this model is effective overall, although there are some challenges in classifying certain types of Wayang, particularly Krucil.

*3.4. YOLOv5-cls model applied to Wayang*

These are the testing results and confusion metrics for the YOLOv5-cls model applied to Wayang classification.

**Table 3** YOLOv5-cls model applied to Wayang.

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Beber | 0 | 0 | 0 |
| Gedog | 0 | 0 | 0 |
| Golek | 0.57 | 1.0 | 0.7272 |
| Krucil | 0.66 | 0.66 | 0.6666 |
| Kulit | 0.4 | 0.8 | 0.5333 |
| Suluh | 0.5 | 0.25 | 0.3333 |
| Overall | | | 0.4782 |

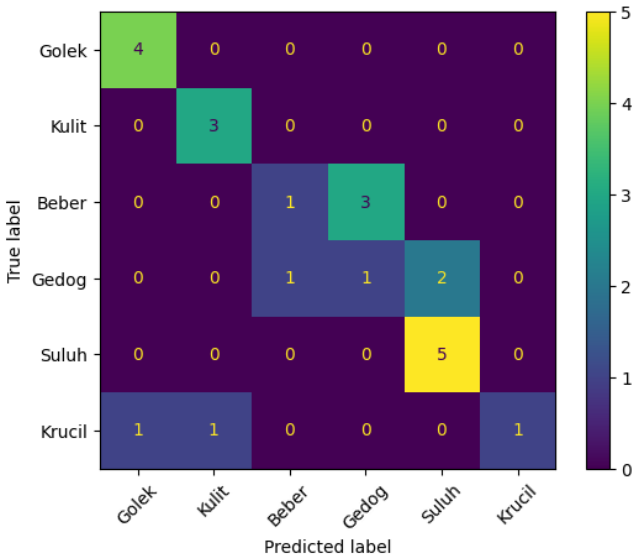


**Figure 4** Confusion Metrics of YOLOv5-cls.

The YOLOv5-cls model is not very good at predicting certain types of Wayang. It struggles with Gedog and Beber Wayang, while Krucil and Golek are recognized relatively well but not perfectly. Suluh Wayang also has unsatisfactory results, and Kulit Wayang is frequently misidentified. The confusion metrics indicates that the model is not very good at predicting certain types of Wayang, particularly Gedog and Beber Wayang.

*3.5. YOLOv8-cls model applied to Wayang classification*

These are the testing results and confusion metrics for the YOLOv8-cls model applied to Wayang classification.

**Table 4** YOLOv8-cls model applied to Wayang classification.

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Beber | 1.0 | 1.0 | 1.0000 |
| Gedog | 0.75 | 1.0 | 0.8571 |
| Golek | 1.0 | 1.0 | 1.0000 |
| Krucil | 1.0 | 0.5 | 0.6666 |
| Kulit | 0.83 | 1.0 | 0.9090 |
| Suluh | 1.0 | 1.0 | 1.0000 |
| Overall | | | 0.9130 |

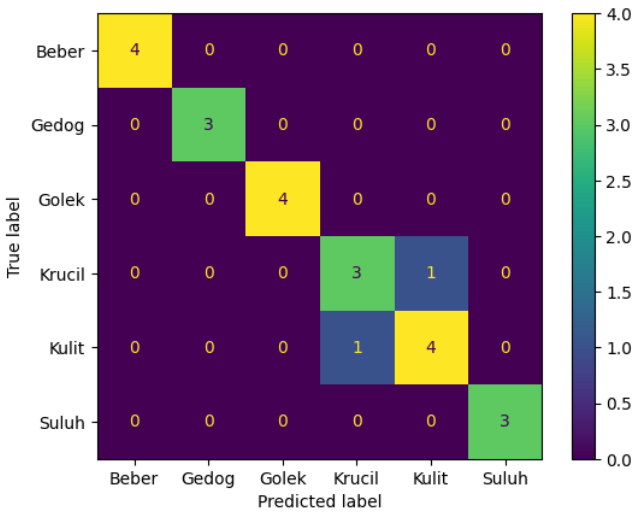


**Figure 5** Confusion Metrics of YOLOv8-cls

The YOLOv8-cls model performed well overall, excelling in classifying Golek, Suluh, and Beber Wayang with perfect precision, recall, and F1-Score. Gedog Wayang shows good results with minor errors, but Krucil Wayang has less satisfactory performance with low parameter values. The confusion metrics highlights that although the model works fairly well, there are still some issues in classifying Wayang with similar characteristics, such as Kulit and Krucil.

*3.6. ViT model applied to Wayang classification*

These are the testing results and confusion metrics for the ViT model applied to Wayang classification.

**Table 5** ViT model applied to Wayang classification

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Beber | 1.0 | 1.0 | 1.000 |
| Gedog | 0.75 | 1.0 | 0.8571 |
| Golek | 1.0 | 1.0 | 1.000 |
| Krucil | 0.75 | 0.75 | 0.7500 |
| Kulit | 1.0 | 0.8 | 0.8888 |
| Suluh | 1.0 | 1.0 | 1.000 |
| Overall | | | 0.9130 |

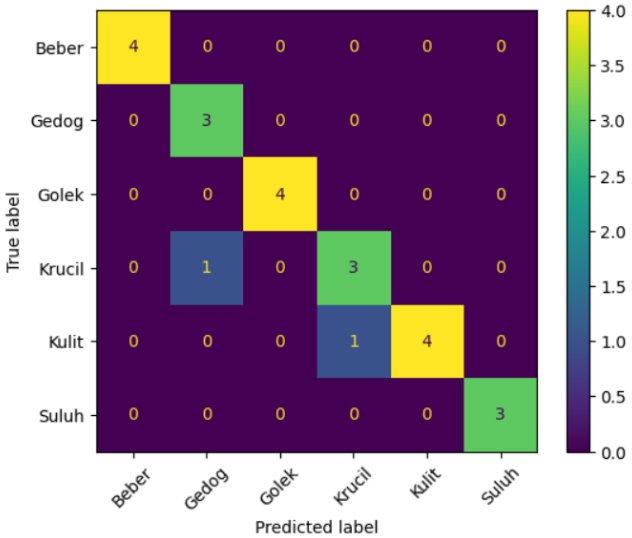


**Figure 6** Confusion Metrics of ViT

The ViT model performs well across most Wayang types, with excellent classification results and perfect parameters for nearly all categories. Krucil and Kulit Wayang are also recognized effectively, despite their similarities. The confusion metrics indicates that the model has very satisfactory performance, although there are slight errors in the classification of Krucil and Kulit Wayang.

## 4. Conclusions

This research successfully developed a Wayang image classification model using several deep learning architectures, namely ResNet34, YOLOv5-cls, YOLOv8-cls, and Vision Transformer (ViT). The test results showed that the ViT and YOLOv8-cls models had the highest accuracy in classifying Wayang types, with ViT achieving an accuracy rate of 91.3%. ViT demonstrated advantages in efficiency and flexibility, being able to handle various image sizes and datasets well. Additionally, ViT excelled in capturing relationships between image parts through the self-attention mechanism.

However, ViT also has some drawbacks. This model requires more computational resources compared to traditional CNN models like ResNet34. Furthermore, ViT tends to be more complex in the training and parameter tuning process, which can be challenging for researchers with limited resources. Despite these challenges, ViT remains a promising choice for complex image recognition tasks, including efforts to preserve Wayang culture.

**Author Contributions:** The concept of this paper comes from Andreas dan Muharrom. Gredy and Theressa are who writes this paper. Moreover, Zaky and Reno help to analyst and running the experiment.

**Funding:**This research received no external funding.

**Data Availability Statement:** The dataset in this research was sourced from Kaggle and later moved to Roboflow for easier access, available on https://universe.roboflow.com/Wayang-cumb2/Wayang-classification.

**Acknowledgment:**Thanks to everyone who always be there for us.

**Conflict of Interest:** The authors declare no conflict of interest

**References**

[1] Putrajip, M. Y., & Retnowati, T. H. (2019). Nilai Edukatif Wayang Ukur Panakawan Karya Sigit Sukasman Dan Implementasinya Pada Pembelajaran Seni Budaya Kelas X SMA. Lumbung Pustaka UNY.

[2] Kusbiyanto, M. (2015). Upaya Mencegah Hilangnya Wayang Kulit Sebagai Ekspresi Budaya Warisan Budaya Bangsa. Jurnal Hukum dan Pembangunan, 45(4), 589-602.

[3] Alfaqi, M. Z. (2022). Eksistensi dan peroblematika pelestarian Wayang kulit pada generasi muda Kec. Ringinrejo Kab. Kediri. Jurnal Praksis dan Dedikasi (JPDS), 5(2), 119-128.

[4] Mulyono, S. (1979). Simbolisme dan Mistikisme dalam Wayang. Jakarta. Gunung Agung.

[5] Aizid, R. (2013). Atlas Pintar Dunia Wayang. Yogyakrta: Diva Press.

[6] Sunardi, Suwarno, B., & Pujiono, B. (2014). Revitalisasi dan inovasi Wayang gedog. ISI Press Surakarta.

[7] Afifah, N. (2019). Makna simbolik Wayang golek jawa barat. Jakarta: Fakultas Ushuluddin dan Filsafat UIN Syarif Hidayatullah.

[8] Widagdo, J. (2015). Struktur Wajah, Aksesoris Serta Pakaian Wayang Golek Menak. Jurnal DISPROTEK, 6(1), 95-102.

[9] Kusumaning Tiyas, S. (2022). Media Wayang Kulit dalam Pembelajaran Bahasa Jawa di Sekolah Dasar. Kalam Cendekia: Jurnal Ilmiah Kependidikan, 10(2).

[10] Siskawati, A., & Alrianingrum, S. (2018). Wayang Suluh Madiun Tahun 1947-1965. AVATARA, e-Journal Pendidikan Sejarah, 6(2), 1-8.

[11] Dradjat, R. P., Darmayanti, T. E., & Isfiaty, T. (2022). Membaca visual Wayang beber sebagai ide perancangan ruang. Visual Heritage: Jurnal Kreasi Seni dan Budaya, 4(3), 309-317.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv.org, Dec. 10, 2015. https://arxiv.org/abs/1512.03385

[13] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved Residual Networks for Image and Video Recognition," in 2020 25th International Conference on Pattern Recognition (ICPR), Jan. 2021. Accessed: Jul. 22, 2024. [Online]. Available: http://dx.doi.org/10.1109/icpr48806.2021.9412193

[14] A. Ridhovan and A. Suharso, "PENERAPAN METODE RESIDUAL NETWORK (RESNET) DALAM KLASIFIKASI PENYAKIT PADA DAUN GANDUM," JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), vol. 7, no. 1, pp. 58–65, Feb. 2022, doi: 10.29100/jipi.v7i1.2410.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," arXiv.org, Jun. 08, 2015.

https://arxiv.org/abs/1506.02640.

[16] Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2021). "A Review of Yolo Algorithm Developments". Procedia Computer Science, 199, 1066-1073. https://doi.org/10.1016/j.procs.2022.01.135.

[17] Ultralytics, "YOLOv5," Ultralytics YOLO Docs, Nov 22, 2022. [Online]. Available: https://github.com/ultralytics/yolov5. [Accessed: Jul. 23, 2024].

[18] Liu, H., Sun, F., Gu, J., & Deng, L. (2022). "SF-YOLOv5: A Lightweight Small Object Detection Algorithm Based on Improved Feature Fusion Mode". MDPI: Sensors, 22(15). https://doi.org/10.3390/s22155817.

[19] R . Dwiyanto, D. W. . Widodo, and P. . Kasih, "Implementation of You Only Look Once (YOLOv5) Method for Vehicle Classification in Tulungagung Regency CCTV". Seminar Nasional Inovasi Teknologi (SEMNAS INOTEK), 3(3), vol. 6, no. 3, pp. 102-104, Nov. 2022.

[20] Ultralytics, "YOLOv8," Ultralytics YOLO Docs, Nov. 12, 2023. Accessed: Jul. 25, 2024. [Online]. Available: https://docs.ultralytics.com/models/yolov8/.

[21] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS", Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1680-1716, Nov. 2023, doi: 10.3390/make5040083.

[22] E. Soylu and T. Soylu, "A performance comparison of YOLOv8 models for traffic sign detection in the Robotaxi-full scale autonomous vehicle competition," Multimedia Tools and Applications, vol. 83, no. 8, pp. 25005-25035, Aug. 2023, doi: 10.1007/s11042-023-16451-1.

[23] K. Han et al., "A Survey on Vision Transformer," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87-110, 1 Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[24] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations, Jun. 03, 2021. https://iclr.cc/virtual/2021/poster/3013 (accessed Jul. 18, 2024).

[25] A. Pangestu, B. Purnama, and R. Risnandar, "Vision Transformer untuk Klasifikasi Kematangan Pisang," Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 11, no. 1, pp. 75–84, Feb. 2024, doi: 10.25126/jtiik.20241117389.

[26] Asmara, I. G. N. B. P., Kesiman, M. W. A., & Indrawan, G. (2023). Balinese Shadow Wayang Characters Detection in the Wayang Peteng Performance Using the YOLOv5 Algorithm. Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI, 12(3), 388-397

[27] Mutawalli, L., Zaen, M. T. A., & Yuliadi. (2023). Komparasi CNN dengan ResNet Untuk Klasifikasi Paling Akurat Tingkat Keganasan Diabetes Berdasarkan Citra Retinopathy. Journal of Computer System and Informatics (JoSYC), 4(3), 522-529

[28] Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P., Korski, K., & Gaire, F. (2022). A comparative study between vision transformers and CNNs in digital pathology. arXiv preprint

[29] Kim, S., & Lee, S. (2024). YOLO-Based Damage Detection with StyleGAN3 Data Augmentation for Parcel Information-Recognition System. Computational Materials Science, 052070

[30] Pande, S. D., & Agarwal, R. (2024). Multi-class kidney abnormalities detecting novel system through computed tomography. IEEE Access, 12, 21147-21159

[31] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations, Jun. 03, 2021. https://iclr.cc/virtual/2021/poster/3013 (accessed Jul. 18, 2024)*Phys. Conf. Ser.*, vol. 1544, no. 1, Jun. 2020, doi: 10.1088/1742-6596/1544/1/012003.