International Journal of Advanced Smart Convergence Vol.13 No.2 195-204 (2024)
http://dx.doi.org/10.7236/IJASC.2024.13.2.195

IJASC 24-2-24

Evaluating Chest Abnormalities Detection: YOLOv7 and Detection Transformer with CycleGAN Data Augmentation

Yoshua Kaleb Purwanto[1], Suk-Ho Lee[2], Dae-Ki Kang[2]

[1]*Master Student, Department of Computer Engineering, Dongseo University, Busan, Korea*
[2]*Professor, Department of Computer Engineering, Dongseo University, Busan, Korea*
*yoshuakaleb049@gmail.com, petrasuk@gmail.com, *dkkang@dongseo.ac.kr*

Abstract

In this paper, we investigate the comparative performance of two leading object detection architectures, YOLOv7 and Detection Transformer (DETR), across varying levels of data augmentation using CycleGAN. Our experiments focus on chest scan images within the context of biomedical informatics, specifically targeting the detection of abnormalities. The study reveals that YOLOv7 consistently outperforms DETR across all levels of augmented data, maintaining better performance even with 75% augmented data. Additionally, YOLOv7 demonstrates significantly faster convergence, requiring approximately 30 epochs compared to DETR's 300 epochs. These findings underscore the superiority of YOLOv7 for object detection tasks, especially in scenarios with limited data and when rapid convergence is essential. Our results provide valuable insights for researchers and practitioners in the field of computer vision, highlighting the effectiveness of YOLOv7 and the importance of data augmentation in improving model performance and efficiency.

Keywords: *Object detection; Computer Vision; YOLOv7; Detection Transformer; Medical Imaging; Data Augmentation; CycleGAN; Generative Adversarial Networks; Performance Evaluation*

1. Introduction

In the ever-evolving landscape of computer vision, object detection stands as a cornerstone task, essential for myriad applications ranging from autonomous vehicles to medical diagnosis. As of 2024, significant strides have been made in the field, driven by the relentless pursuit of accuracy, efficiency, and adaptability to diverse real-world scenarios.

Amidst these advancements, one of the most prominent families of object detection models is the You Only Look Once (YOLO) series. YOLO revolutionized the field with its unified approach, performing both object localization and classification in a single pass through the neural network. This real-time capability made YOLO models invaluable for applications requiring rapid decision-making, such as video surveillance and robotics [1][2][3][4].

Manuscript Received: May. 6, 2024 / Revised: May. 16, 2024 / Accepted: May. 21, 2024
Corresponding Author: dkkang@dongseo.ac.kr
Tel: +82-51-320-1724, Fax: +82-51-327-8955
Professor, Department of Computer Engineering, Dongseo University, Busan, Korea

Meanwhile, the introduction of transformer architectures, initially developed for natural language processing tasks, has sparked a paradigm shift in computer vision. Transformers have demonstrated remarkable success in various vision tasks, owing to their ability to capture global dependencies and long-range interactions within the input data. In the context of object detection, models such as the Detection Transformer (DETR) have emerged, leveraging transformer architecture to directly predict object bounding boxes and class labels without relying on traditional anchor-based approaches [5].

Data augmentation has long been recognized as a critical component in training robust and generalizable deep learning models. By artificially augmenting the training data with diverse variations, models can better adapt to variations in the input data encountered during inference. Recent advancements in data augmentation techniques, such as generative adversarial networks (GANs), have further expanded the repertoire of augmentation strategies. CycleGAN, a notable example, enables unpaired image-to-image translation, allowing for realistic transformations between different domains [10]. By leveraging CycleGAN, researchers can generate augmented data with minimal manual effort, facilitating the training of more robust object detection models.

In this paper, we delve into the realm of object detection, exploring the comparative performance of two leading architectures: YOLOv7 and Detection Transformer. Specifically, we investigate their efficacy across varying levels of data augmentation using CycleGAN, focusing on chest scan images in the context of biomedical informatics. Through rigorous experimentation and analysis, we aim to provide insights into the strengths and limitations of each approach and their implications for real-world applications in medical imaging and beyond.

2. Related Work

2.1 You Only Look Once (YOLO)

The You Only Look Once (YOLO) series has undergone several iterations, each introducing improvements in speed, accuracy, and feature representation. YOLOv1 pioneered the concept of real-time object detection by framing the task as a single regression problem, achieving impressive results albeit with limitations in detecting small objects [2]. Subsequent versions, such as YOLOv3 and YOLOv4, addressed these limitations by introducing architectural enhancements and multi-scale detection strategies, significantly improving performance across various datasets and object sizes [3][4].

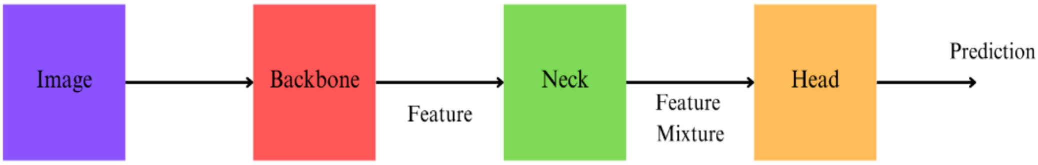

Figure 1. YOLO Architecture Streamline

The main point of object detection algorithm is fast and strong on detecting a feature of a certain object [1]. YOLOv7, short for "You Only Look Once version 7," represents a significant advancement in object detection architecture within the realm of computer vision. Developed as an evolution of previous YOLO models, YOLOv7 introduces several key innovations that enhance its performance and efficiency. One notable feature is its streamlined architecture, which optimizes the balance between accuracy and speed, making it well-suited for real-time applications. This efficiency is achieved through a carefully designed network structure that reduces computational complexity while preserving the model's ability to accurately detect objects across various scales and categories.

In conducting object detection using YOLOv7, the algorithm demonstrates rapid training and high accuracy. According to the research findings, YOLOv7 can achieve optimal training within just 30 epochs [1]. This is attributed to the optimization of the algorithm itself, which includes model scaling for concatenation-based models. Furthermore, YOLOv7 is equipped with planned re-parameterized convolution and coarse for auxiliary and fine for lead loss [1]. These optimizations enable the algorithm to swiftly and effectively reach its optimal convergence point.

Another distinctive aspect of YOLOv7 is its flexibility and adaptability to different deployment scenarios. The architecture allows for seamless integration with diverse hardware platforms, enabling deployment on both resource-constrained edge devices and high-performance computing systems. This versatility makes YOLOv7 an attractive choice for a wide range of applications, including autonomous vehicles, surveillance systems, and robotics, where real-time object detection is critical. Additionally, YOLOv7 incorporates state-of-the-art techniques in deep learning, such as attention mechanisms and feature pyramid networks, further enhancing its capability to detect objects with high accuracy and robustness in complex scenes.

2.2 Detection Transformer (DETR)

Originally designed for natural language processing tasks, transformer architectures have found remarkable success in computer vision applications [7]. The transformer's self-attention mechanism enables capturing long-range dependencies within the input data, making it well-suited for tasks requiring global context understanding, such as image classification, object detection, and segmentation [6][8].

In the domain of object detection, the Detection Transformer (DETR) stands out as a pioneering model that replaces traditional convolutional layers with transformer blocks. By directly predicting object bounding boxes and class labels without the need for anchor boxes or region proposal networks, DETR offers a streamlined approach to object detection [5]. This paradigm shift has led to significant improvements in accuracy and efficiency, positioning transformer-based models as formidable competitors to traditional convolutional neural network (CNN) architectures.

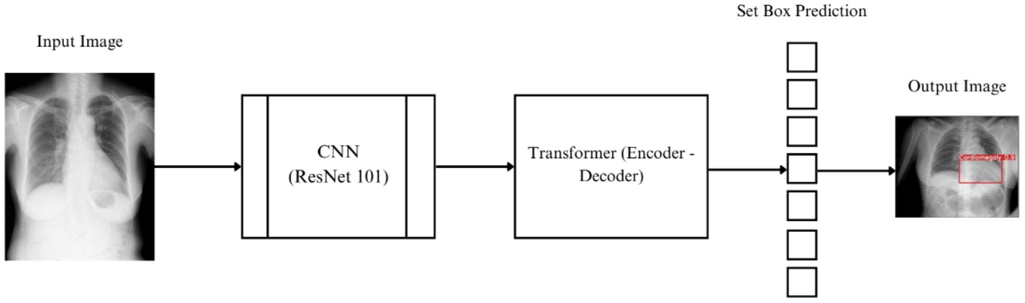

Figure 2. Detection Transformer

As in figure 2, one of the standout abilities of DETR is its capacity to handle variable numbers of objects in an image without the need for predefined anchors or bounding boxes [5]. This flexibility allows DETR to accurately detect objects across a wide range of scales and aspect ratios, making it particularly well-suited for scenarios with complex scenes or densely packed objects. Furthermore, DETR exhibits remarkable generalization capabilities, demonstrating robust performance even on datasets with diverse object categories and challenging backgrounds. Its ability to learn rich spatial relationships and contextual information from input images enables DETR to achieve state-of-the-art results in object detection tasks while offering a more streamlined and interpretable architecture compared to traditional CNN-based approaches.

2.3 CycleGAN

Generative adversarial networks (GANs) have revolutionized the field of image generation, enabling the creation of realistic synthetic images from random noise. CycleGAN extends this concept to the domain of unpaired image-to-image translation, allowing for the transformation of images from one domain to another without the need for paired training data [10].

The core idea behind CycleGAN is the cycle consistency loss, which enforces the translated images to be consistent when translated back to the original domain. This ensures that the transformation process captures meaningful semantic features while preserving essential characteristics of the input images. CycleGAN has been widely adopted for various tasks, including style transfer, image super-resolution, and domain adaptation, making it a versatile tool for data augmentation and image manipulation tasks.

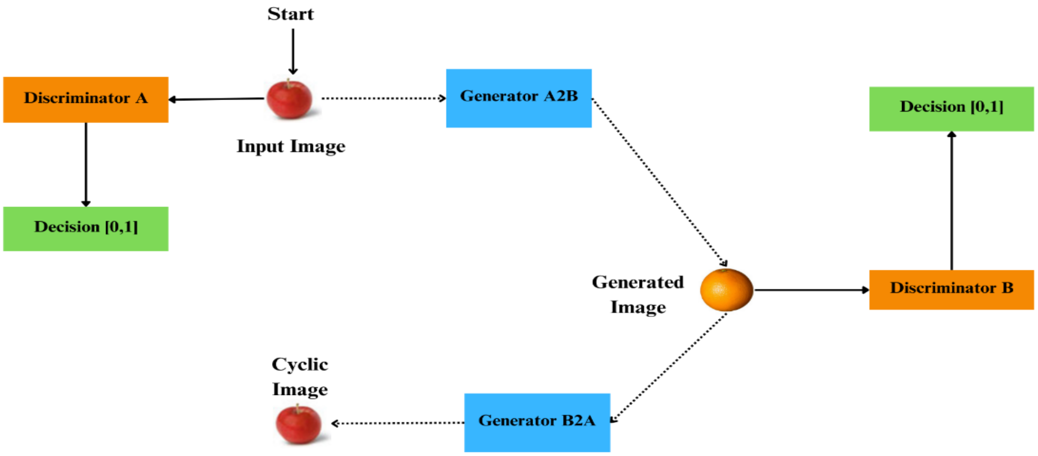

Figure 3. CycleGAN Network

CycleGAN, a variant of Generative Adversarial Networks (GANs), is renowned for its ability to perform unsupervised image-to-image translation between two domains without requiring paired training data. Its architecture consists of two main components: a generator and a discriminator as in figure 3. The generator learns to map images from one domain to another, while the discriminator aims to distinguish between translated images and real images from the target domain [11][12]. What sets CycleGAN apart is its incorporation of cycle-consistency loss, which enforces the condition that translating an image from one domain to another and back should result in the original image. This constraint encourages the model to learn meaningful mappings between domains while preserving essential visual attributes, leading to high-quality translations even in the absence of direct supervision.

The standout ability of CycleGAN lies in its capacity to learn domain mappings in an unsupervised manner, making it highly versatile for various image translation tasks. Whether it's transforming photographs into artistic renditions, converting images between different seasons or landscapes, or even altering attributes like style or age, CycleGAN excels at capturing complex relationships between domains and producing realistic results. Its ability to learn without paired training data offers significant advantages in scenarios where collecting such data may be challenging or impractical. Additionally, CycleGAN's architecture is modular and adaptable, allowing for easy extension and customization to suit specific applications and datasets. This combination of versatility, effectiveness, and ease of use has cemented CycleGAN as a powerful tool for image manipulation and synthesis in both research and practical applications.

2.4 Dataset: VinDr-CSR

In recent years, the development of machine learning algorithms for the detection and localization of chest abnormalities in X-ray scans has faced significant challenges due to the limited availability of annotated datasets. Existing chest X-ray datasets typically provide labels for findings without specifying their precise locations on the radiographs, hindering the progress of automated diagnosis systems requiring detailed annotations for accurate localization of abnormalities.

To address this limitation, a large-scale collection of chest X-ray images has been meticulously annotated by experienced radiologists, providing researchers with access to a valuable dataset. Consisting of 3076 images, the data is manually annotated with 14 local labels as in Table 1. Each scan in the training set is independently labeled by three radiologists, while a consensus of five radiologists labeled each scan in the test set [9]. This meticulous annotation procedure enhances the dataset's quality and enables robust evaluation of machine learning models' performance on detecting and localizing chest abnormalities.

Table 1. Dataset Classification

ID	Label
0	Aortic enlargement
1	Atelectasis
2	Calcification
3	Cardiomegaly
4	*Consolidation*
5	ILD
6	Infiltration
7	Lung Opacity
8	Nodule/Mass
9	Other Lesion
10	Pleural Effusion
11	Pleural Thickening
12	Pneumothoras
13	Pulmonary Fibrosis

3. Experiment Setup and Methodology

3.1 Setup Environment

Table 2. System Environment

Computer Environment	Software frame and Library
GPU: RTX 4090	CUDA Version 12.4
CPU: AMD Ryzen 9 16-Core	Cudnn 8.9
RAM: 32 GB DDR5	Pytorch 2.2.2
OS: Windows 11	

For this experiment, we used a computational environment optimized for high-performance deep learning tasks, providing a robust platform for cutting-edge research and development. Anchored by an RTX 4090 GPU with CUDA version 12.4 and CuDNN 8.9, the system utilizes NVIDIA's flagship graphics card for accelerated training and inference tasks. Complemented by an AMD Ryzen 9 16-Core CPU and 32 GB of DDR5 RAM, it efficiently handles preprocessing, data augmentation, and post-processing operations. Operating on Windows 11, the system leverages PyTorch 2.2.2, a flexible and scalable deep learning framework, enabling researchers and developers to implement advanced neural network architectures and experiment with optimization algorithms. Combined, these hardware and software components create a versatile computational environment for deep learning research and application development.

3.2 Data Augmentation Scenario

Using CycleGAN as the data augmentation algorithm, chest scan images are provided as input and undergo translation operations from the medical imaging domain to a domain with either higher-contrast colors, different color patterns, or color shifts. The differences between the two domains create changes in color, texture, and contrast, which can aid in scenarios where there is limited data or insufficient annotations to achieve good detection performance. However, excessive data augmentation can lead to overfitting. Therefore, data augmentation must be carefully controlled to prevent the network from performing poorly on test data.

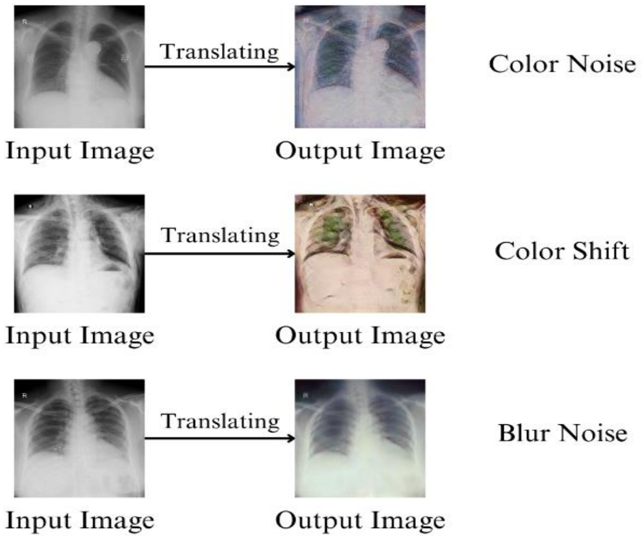

Figure 4. Data Augmentation

With the created augmented data, augmentation is performed by dividing it into 0%, 25%, 50%, and 75% augmentation levels. The evaluation scenario for the algorithm without using data augmentation involves 4394 images divided into 70% for training, 20% for validation, and 10% for testing. This partitioning results in training, validation, and test datasets containing 3076, 879, and 439 images, respectively. In cases where data augmentation is applied, a percentage of the original images is removed according to the chosen augmentation level. Subsequently, the original images are selected for augmentation until they match the dataset size. This approach ensures that the evaluation scenario remains aligned with the actual dataset size, enabling meaningful comparisons.

Table 3. Data Augmentation Division

Percentage	Augmented Data Total	Original Image Total
0%	0	3076
25%	769	2307
50%	1538	1538
75%	2307	769

3.3 Chest Abnormalities Detection

In the scenario of chest abnormality detection, evaluation is conducted by training the algorithms under original performance settings. YOLOv7 and DETR detectors are employed to assess the capabilities of each detector under different data augmentation percentage scenarios. The evaluation involves training the models for 100 epochs using pre-trained models provided by the researchers who developed the algorithms. The pre-training is conducted with optimal and effective parameter settings.

In detecting abnormalities, each algorithm is executed using its respective pre-trained model. This is done to ensure that each algorithm and parameter used are optimal. The comparison is based on the training time required, the number of parameters needed, and the Average Precision (AP) value from the validation results. The comparative analysis of the two algorithms is determined according to the amount of data augmentation in the training scenario. The baseline is established as the training results without using data augmentation by each algorithm.

4. Result and Discussion

Result. YOLOv7 consistently outperforms DETR across all levels of augmented data. Even with 75% augmented data, YOLOv7 maintains better performance compared to DETR. Both models exhibit a decrease in performance as the amount of augmented data increases, but YOLOv7's performance degrades more gracefully compared to DETR. Despite the decrease in performance, YOLOv7 remains more robust to the amount of augmented data compared to DETR.

Table 4. YOLOv7 Validation Result

Augmented Data Percentage	AP_{50}	AP	Training Time	#Parameter
0%	0.4	0.2	4.1 Hours	37.26 M
25%	0.385	0.188	4.1 Hours	37.26 M
50%	0.366	0.169	4.1 Hours	37.26 M
75%	0.08	0.04	4.1 Hours	37.26 M

Table 5. DETR Validation Result

Augmented Data Percentage	AP_{50}	AP	Training Time	#Parameter
0%	0.293	0.135	6 Hours	41.26 M
25%	0.247	0.091	6 Hours	41.26 M
50%	0.159	0.071	6 Hours	41.26 M
75%	0.071	0.038	6 Hours	41.26 M

Analysis. YOLOv7 requires significantly fewer epochs compared to DETR to reach convergence or the optimal point for weight optimization. Typically, YOLOv7 reaches convergence within approximately 30 epochs, whereas DETR requires around 300 epochs to converge effectively. This stark difference in convergence speed can be attributed to the architectural differences between the two models. YOLO's architecture, particularly its one-stage object detection approach, allows for faster convergence due to its simplicity and efficiency in processing. On the other hand, DETR's transformer-based architecture, while powerful and capable of capturing global context effectively, requires more epochs to converge due to its complexity and the need for longer training to effectively learn the parameters. Therefore, YOLOv7 offers a significant advantage in terms of training speed and efficiency compared to DETR.

The test results demonstrate the impact of data augmentation on the performance of YOLOv7 and DETR object detection models. While data augmentation enhances the models' performance initially by providing additional diverse training samples, there is a trade-off as the amount of augmented data increases. For YOLOv7, with 0% augmented data, the model achieves an AP of 0.4, which gradually decreases to 0.08 with 75% augmented data. Similarly, for DETR, the AP decreases from 0.293 with 0% augmented data to 0.071 with 75% augmented data. Although data augmentation improves the models' generalization capabilities initially, excessive augmentation can lead to overfitting, resulting in decreased performance on unseen data. Thus, the choice of the amount of augmented data requires a balance between improving model performance and preventing overfitting, as indicated by the diminishing AP values with increasing augmented data percentages.

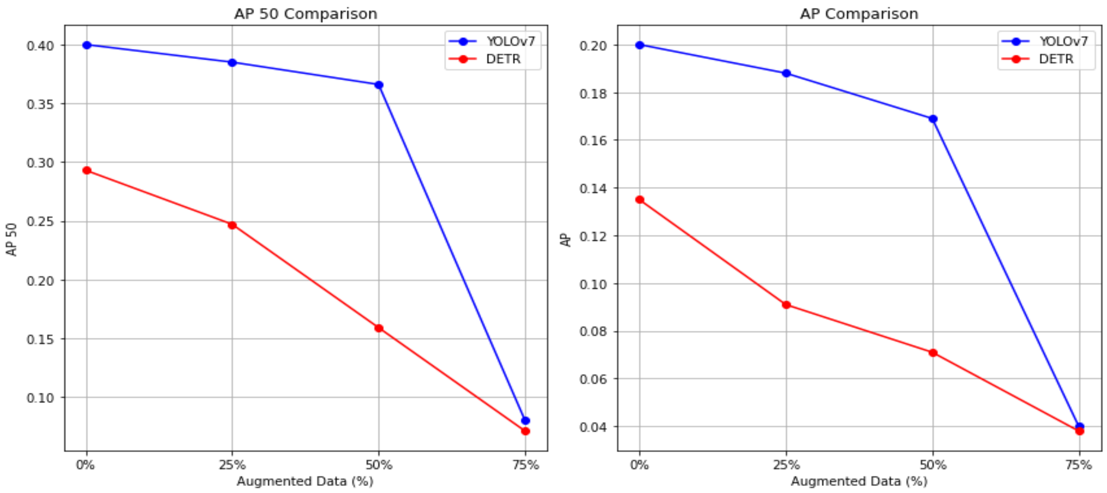

Figure 5. Average Precision Graph

5. Conclusion

In this study, we compared the performance of two leading object detection architectures, YOLOv7 and Detection Transformer (DETR), across varying levels of data augmentation using CycleGAN. Our experiments on chest scan images within the context of biomedical informatics revealed that YOLOv7 consistently outperforms DETR across all levels of augmented data. Even with 75% augmented data, YOLOv7 maintains better performance compared to DETR, with a more graceful degradation in performance as the amount of augmented data increases. Additionally, YOLOv7 demonstrates significantly faster convergence, requiring approximately 30 epochs compared to DETR's 300 epochs. These findings underscore the superiority of YOLOv7 for object detection tasks, especially in scenarios with limited data and when rapid convergence is essential. Further research is warranted to explore the generalization capabilities of these models on diverse datasets and to optimize training parameters for improved performance and efficiency in real-world applications.

Acknowledgement

This work was supported by Dongseo University, "Dongseo Cluster Project" Research Fund of 2023 (DSU-20230004).

References

[1] A. Bochkovskiy, C. Wang, H. M. Liao, and R. Girshick, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Apr. 2021.
DOI: https://doi.org/10.48550/arXiv.2207.02696

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
DOI: https://doi.org/10.48550/arXiv.1506.02640

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “YOLOv3: An Incremental Improvement,” Apr. 2018.
DOI: https://doi.org/10.48550/arXiv.1804.02767

[4] A. Bochkovskiy, C. Wang, and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020.
DOI: https://doi.org/10.48550/arXiv.2004.10934

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in European Conference on Computer Vision (ECCV), 2020, pp. 213–229. Springer.
DOI: https://doi.org/10.48550/arXiv.2005.12872

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems (NeurIPS), 2017.
DOI: https://doi.org/10.48550/arXiv.1706.03762

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Oct. 2020.
DOI: https://doi.org/10.48550/arXiv.2010.11929

[8] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision Tranformer Adapter for Dense Predictions," In Proceedings of the 9th International Conference on Learning Representations (ICLR), Feb. 2023.
DOI: https://doi.org/10.48550/arXiv.2205.08534

[9] H. Q. Nguyen et al., "VinDr-CXR: An Open Dataset of Chest X-rays with Radiologist’s Annotations," Jan. 2022.
DOI: https://doi.org/10.48550/arXiv.2012.15029

[10] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
DOI: https://doi.org/10.48550/arXiv.1703.10593

[11] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in European Conference on Computer Vision (ECCV), 2016.
DOI: https://doi.org/10.48550/arXiv.1603.08511

[12] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[13] M. Tan and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
DOI: https://doi.org/10.48550/arXiv.1911.09070

[14] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "Learning non-maximum suppression," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
DOI: https://doi.org/10.48550/arXiv.1705.02950

[15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proceedings of the 3rd International Conference on Learning Representations (ICLR), Sep. 2014.
https://doi.org/10.48550/arXiv.1409.0473