International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 5
ISSN: 2088-5334

Robust Estimation of Crowd Density Using Vision Transformers

Chuho Yi [a], Jungwon Cho [b,*]

[a] *Department of AI Convergence, Hanyang Women's University, Seoul, Republic of Korea*
[b] *Department of Computer Education, Jeju National University, Jeju, Republic of Korea*
*Corresponding author: *jwcho@jejunu.ac.kr*

***Abstract*— Estimating and predicting crowd density at large events or during disasters is of paramount importance for enhancing emergency management systems. This includes planning effective evacuation routes, optimizing rescue operations, and ensuring efficient deployment of emergency services. Traditionally, surveillance systems that rely on cameras have been employed to monitor crowd movements. However, accurately estimating crowd density using such systems presents several challenges. These challenges stem primarily from the interaction between large crowds and the limitations of two-dimensional cameras in capturing the full scope of three-dimensional spaces. Optical distortions, environmental factors, and variations in camera angles further complicate the task, making accurate estimations difficult to achieve. To address these challenges, this paper introduces a robust method for calculating crowd density that leverages advanced vision transformers. By combining the output of these transformers with a two-stage neural network, the method effectively mitigates the limitations of traditional approaches. One of the key advantages of the proposed system is its robustness, which allows it to perform well across different camera specifications, installation locations, and image aspect ratios. The method applies and evaluates various deep learning techniques, introducing improvements to existing network structures that are better suited for the problem at hand. Extensive experimental verification demonstrates that the proposed method consistently produces accurate crowd density estimates, even in diverse and complex crowd environments. This robust performance underscores its potential for improving emergency management and crowd control in real-world situations.**

***Keywords*— Estimation of crowd density; vision transformer; data augmentation as feature manipulation; Deep Crowd Density Network (DCDN); Convolutional Crowd Density Network (CCDN).**

Manuscript received 15 Dec. 2023; revised 29 Mar. 2024; accepted 11 Jun. 2024. Date of publication 31 Oct. 2024.

I. INTRODUCTION

Estimating crowd density is important in various situations, such as for public safety, urban planning, and traffic management. Accurately determining the density of crowds, especially at significant events or in public transport facilities, is essential for enhancing safety and the efficiency of resource allocation. Against this backdrop, computer vision technology has become crucial for automatically estimating crowd density from video sources like closed-circuit television, enabling real-time monitoring and analysis [1], [2].

Most vision-based crowd density estimation methods have been optimized for static camera setups and consistent environmental conditions. In real-world settings, however, cameras' specifications and installation positions, capture angles, and image ratios can vary significantly and greatly affect the accuracy of density estimates. Traditional methods are vulnerable to environmental changes, especially when processing images taken from multiple locations and angles, often resulting in significant accuracy issues [3].

Anticipating abnormal events or emergencies enables the mobilization of necessary resources in response. This allows the efficient deployment of safety personnel, such as police or administrators, and crowds can be safely guided or managed using public address systems. Pre-planned response strategies can enhance on-site management during emergencies and help minimize casualties. Therefore, estimating and predicting crowd densities accurately is critical in emergency management.

While monitoring crowd complexities with cameras is generally sound, accurately understanding crowds' three-dimensional spatial positions and interactions can be challenging with two-dimensional cameras. This can hinder a comprehensive understanding of crowd depth and spatial structure. Moreover, factors such as lighting conditions, shadows, reflections, and other optical issues or environmental effects can complicate accurate crowd

detection and analysis using cameras. Processing and analyzing extensive crowd data in real-time requires high efficiency in computing resources and algorithms.

This paper proposes a hybrid network that combines vision transformers and neural network methods to overcome the challenges of estimating crowd density using cameras. Research on crowd density prediction using cameras has evolved with advances in vision technology. Some studies by [4] and [5] proposed methods using computer vision technology to measure crowd density in outdoor locations, considering factors such as lighting changes, attire, and weather variation in images. They used a grey-level dependency matrix, Minkowski fractal dimension, and newly developed translation invariant orthonormal Chebyshev moments to extract image features, classifying them into various densities using a self-organizing map. They suggested the best method for vision-based crowd density measurement by comparing three methods.

Saleh et al. discussed advances in automated systems for crowd density estimation and accounting [6], [7]. They reviewed various methods used in computer vision-based surveillance systems for analyzing and managing crowds. They considered direct (object-based target detection) and indirect (pixel-based, texture-based, and edge point-based analysis) approaches for analyzing crowds. Analyzing crowd dynamics and behavior remains an important research topic in psychology, sociology, public services, safety, and computer vision.

The use of cameras for crowd analysis has undergone many changes with the emergence of deep learning [8], [9], [10], particularly with the advent of the Transformer Model. Islam et al. wrote an extensive introduction to the Transformer Model, which comprehensively surveyed the various uses of transformers in deep learning tasks [11]. Transformers use attention mechanisms to understand contextual relationships within sequential data and perform exceptionally in natural language processing (NLP), computer vision, speech and audio processing, healthcare, and the Internet of Things. They reviewed transformer models proposed from 2017 to 2022, identifying and classifying critical models in five main application areas: NLP, computer vision, multimodality, audio and speech processing, and signal processing. They analyzed the impact of transformer-based models in each region and discussed future research directions and possibilities.

Some previous studies proposed methods for applying the Transformer Model to images, notably the Vision Transformer [12], [13]. This method was proposed to identify and focus on areas of interest within images for human recognition or detection tasks. It includes dividing images into multiple patches, flattening each patch into a one-dimensional vector, and feeding these vectors into the transformer architecture. Importantly, positional information between patches is also integrated as model input to preserve spatial information. This model performed well in various vision tasks using embeddings from pre-trained language models on large datasets and applying a transfer learning approach. It has gained attention for its ability to train compelling vision models using extensive datasets and computational resources.

Chen et al. proposed a new method using convolutional neural networks (CNNs) and transformers to count crowds of various densities effectively, introducing the CNN and Transformer Adaptive Selection Network (CTASNet), which automatically selects the appropriate counting branch for areas of varying density [14, 15]. CNNs are effective for target location and counting in low-density regions, while transformers have high reliability in high-density areas. This paper validated the superiority of this method through extensive experiments using four primary crowd-counting datasets. It introduced coherent entropy-based optimal transport loss to reduce the impact of annotation noise.

A multifaceted Attention Network (MAN) has been proposed for counting crowds, effectively handling the considerable variation common in crowd images [16], [17]. This network combines traditional transformer global and local attention to achieve enhanced local spatial relationship encoding. It uses Multifaceted Attention to allocate attention to feature locations and supervise training. It also includes a method focused on the most essential instances during training. Extensive experiments confirmed that this method performed outstandingly on four challenging crowd-counting datasets.

Our proposed method robustly estimates crowd density using vision transformer technology. It effectively captures global and local contexts within images using attention mechanisms, thus implementing a model resilient to various visual changes. This study aimed to achieve more precise density estimation by combining the estimated density information with a two-stage neural network.

First, this paper develops methods that can be applied robustly across various camera specifications, installation positions, and image ratios, introducing data preprocessing and augmentation techniques that consider camera physical constraints. Then, it applies these methods to improve existing network structures to be compatible with various deep learning approaches. This includes experimental approaches to compare the strengths and weaknesses of different deep learning architectures and find the optimal combination.

To validate the performance of the proposed methods, density estimation experiments are conducted in complex crowd situations and various camera environments to verify the method's applicability and efficiency in real-world settings. This research advances crowd density estimation methods and aims for widespread application in real urban environments.

II. Materials and Method

A. Vision Transformer for Crowd Counting

In this paper, we use the method proposed by Lin et al. [16] for counting the number of people in a crowd. This method introduces the Multifaceted Attention Network (MAN) to address the complex problems of crowd counting, proposing ways to enhance the accuracy of crowd counting for various applications, such as video surveillance and traffic management. Its key feature is the introduction of Learnable Region Attention (LRA), which dynamically allocates unique attention regions at each feature location, and Local Attention Regularization (LAR) to supervise the training of LRA, minimizing the deviation in attention. Additionally, Instance Attention is applied to focus on essential instances during exercise, thereby reducing the impact of annotation errors.

This approach has demonstrated excellent performance on crowd-counting datasets such as ShanghaiTech [18], [19], UCF-QNRF [20], JHU++ [21], and NWPU [22]. Technically, the model uses VGG-19 CNN as a backbone to extract initial features fed into a transformer encoder to apply LRA and integrate various loss functions to optimize training. The current study presents a new approach that robustly handles large-scale variation in crowd images, effectively integrates global and local attention mechanisms, and contributes to improving the accuracy of crowd counting and developing models resilient to label noise.

Fig. 1 A target image used for estimating crowd density.

In this paper, we propose a method for robustly detecting individuals within crowds of varying sizes using the vision transformer approach, as illustrated in Fig. 1. Instead of merely counting the number of people to estimate density, we suggest dividing an area into multiple grid-like regions. Subsequently, we use a machine learning training process based on the number of people in each region to compute the complexity of the crowd.

B. Feature Transformation for Using the Output of Vision Transformers as Input

The output from the vision transformer provides feature results for counting the number of people in a crowd. These results need an appropriate transformation formula to calculate the crowd density. As shown in Fig. 2, detection is centered around the people in the image.

Fig. 2 Result using the vision transformer to count the number of people in a crowd.

In Fig. 2, each individual in the crowd is detected successfully despite the complex environment. In the second stage, these results are used as input for the next network on a regional basis. Using the final headcount as input might not yield the desired results if only the number of individuals is used to determine crowd density. Alternatively, using the results from the first stage may lead to undesirable outcomes depending on how the result areas are divided.

Considering real datasets and various surveillance camera settings, the results must be transformed. To apply this robustly across different camera specifications, installation locations, and image ratios, it is essential to preprocess and augment the output from the vision transformer. For this purpose, this paper develops the following formula.

$$M^1 = \sum o(i,j) \tag{1}$$

$$M^2 = \sum R\left(\frac{o(i,j)}{M^1}\right) \tag{2}$$

$$z(i',j') = \frac{M^1}{M^2} R\left(\frac{o(i,j)}{M^1}\right) \tag{3}$$

This formula transforms the output of the vision transformer into features that have the same size and contain the calculated number of people. Equation (1) is the sum of the output values $o(i,j)$, where i and j are the respective row and column positions of the output. Equation (2) sums the transformed values after normalizing the differently sized feature values with the value from equation (1). Here, the function $R\left(\frac{o(i,j)}{M^1}\right)$ uses the resize function from the NumPy library [23, 24], applying bilinear interpolation. Equation (3) is the formula applied to ensure that the resized features contain the number of people where i' and j' are the row and column positions of the transformed features, respectively. In this paper, considering the size and efficiency of the dataset, the features are transformed into 64x64 dimensions.

C. Vision Transformer for Crowd Counting

In typical images, feature values are normalized before input, but this cannot be applied when calculating crowd density in this paper, as it may cause distortion. Additionally, images with fewer people in the training dataset may need more image data, which can result in poorer performance at distinguishing them. To address these issues, this paper modifies the method proposed by [25] and [26], using data augmentation as feature manipulation, and uses the value from equation (1) as input data in the final layer of the network structure. This approach allows the feature values from equation (3) to compensate for the reduced significance of the original crowd numbers with intermediate values.

Fig. 3 outlines the changes in network structure using data augmentation as feature manipulation, as proposed here. Our proposed method describes a network structure that uses data augmentation to manipulate features and goes through two classification stages. The initial data $x(i,j)$ are input through the Vision Transformer Network, which is fixed in a pre-trained state and transforms features without additional training. The transformed features $o(i,j)$ are then input into the first classification network, either a deep neural network (DNN) or a CNN, producing the first set of classification results. We call these methods the Deep Crowd Density Network (DCDN) and Convolutional Crowd Density

Network (CCDN) [27, 28]. A critical aspect of this figure is the application of the data augmentation technique M1. This allows the input feature vectors to receive additional information for learning. These vectors are used as inputs to the second network, either DNN or CNN, which we call Deep Crowd Density Network with Data Augmentation (DCDN+) and Convolutional Crowd Density Network with Data Augmentation (CCDN+). These networks have more refined learning capabilities and, by processing the modified data, they produce more robust and accurate classification results. The results of these experiments will be detailed in the next chapter.

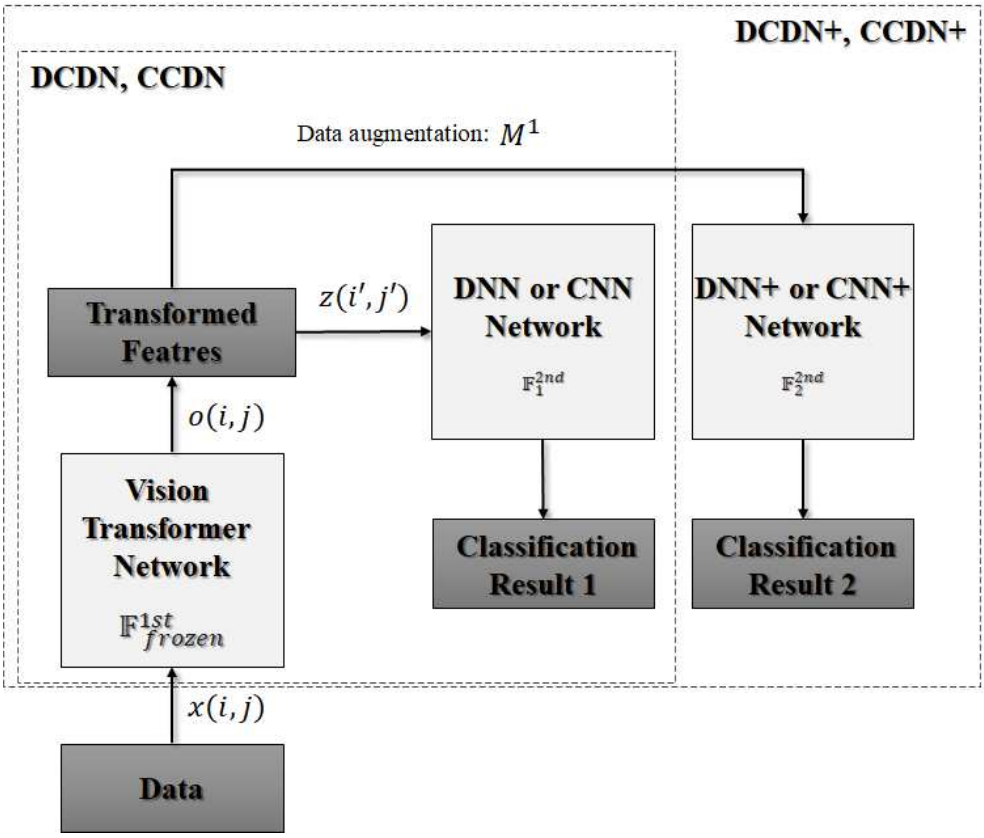

Fig. 3 Proposed changes in network architecture using data augmentation as feature manipulation.

Consequently, this paper proposes a more robust, accurate method for estimating crowd density that integrates vision transformer technology with data augmentation techniques and can adapt to changing environments and various crowd scenarios. This method should perform better than existing methods, mainly when applied to large-scale crowd data.

III. Results and Discussion

The experiments conducted in this paper used an Intel i9 CPU, 32GB of memory, and a Nvidia GeForce RTX 3070 graphics card. The program was implemented using Python, and the training library used was PyTorch [29], [30]. The public dataset used to test the proposed method was the JHU CROWD set [31]. This dataset contains 2,034 images for training and 1,433 images for testing. Fig. 4 shows the images that are categorized in the paper.

The images in Table 1 differ in shooting angle, resolution, and aspect ratio, suggesting that such diversity is essential for estimating crowd density. Images with high crowd densities differ significantly depending on the capture location and angle and also vary in resolution and lighting conditions (day and night). The image on the left was taken indoors in the medium crowd density images. In contrast, the image on the right was taken in a foggy outdoor setting, demonstrating that these environmental factors must be considered when estimating crowd density. The low crowd density images consist of an image taken by the sea with fog and another taken close to individuals.

This study used training and testing data that reflect various conditions to estimate crowd density accurately under these varied environmental conditions. Crowd density was categorized into three levels (high, medium, and low), and a density function appropriate for each category was trained. This is a crucial step to enhance the accuracy of crowd density estimation and strengthen the generalization capabilities of the model.

TABLE I
SAMPLE IMAGES FROM A PUBLIC DATASET FOR ESTIMATING CROWD DENSITY CLASSIFICATION.

Class	Sample images	
High Crowd Density		
Medium Crowd Density		
Low Crowd Density		

Fig. 4 plots the accuracy of the experiments using the test set, based on the methods proposed in the previous chapter, comparing the changes in accuracy while training the DCDN (black), DCDN+ (red), CCDN (blue), and CCDN+ (green) network models.

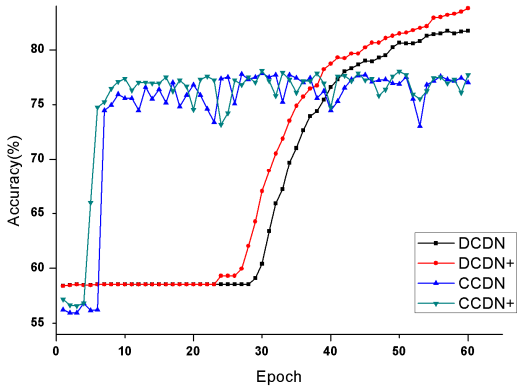

Fig. 4 Experimental results of the method proposed in this paper

Initially, the accuracy of the DCDN model is low, but it improves gradually as training progresses and reaches 78%. By contrast, the accuracy of the DCDN+ model increases rapidly, reaching 81% with training, the best performance among all the models. The CCDN model begins with very low initial accuracy, and its performance improvements are unstable, reaching an accuracy of 75%. The CCDN+ has low initial accuracy but improves significantly after ten epochs, stabilizing at 77% accuracy. These results suggest that the DCDN+ model based on the Vision Transformer has superior learning efficiency and accuracy in crowd density estimation, with the '+' models generally performing better, indicating their further optimization. This analysis provides important insights into how each network structure performs in specific environments.

IV. CONCLUSION

This paper proposes a robust, accurate crowd density estimation method at large event or during disasters. Specifically, we used vision transformers to calculate crowd numbers and combined this with an enhanced two-stage neural network structure to overcome computer vision challenges and environmental effects. The experimental results showed that the proposed DCDN+ model had high accuracy and robustness across various camera specifications, installation positions, and image ratios. We evaluated various deep learning methods and confirmed that this approach is effective in real, complex crowd situations. The method proposed here significantly enhances emergency management and provides essential information on crowd density for planning evacuation routes and emergency services. This research provides important guidelines for designing and implementing crowd management and surveillance systems.

ACKNOWLEDGMENT

This research was supported by the 2023 scientific promotion program funded by Jeju National University.

REFERENCES

[1] B. Li, H. Huang, A. Zhang, P. Liu, and C. Liu, "Approaches on crowd counting and density estimation: a review," *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 853–874, Feb. 2021, doi:10.1007/s10044-021-00959-z.

[2] M. A. Khan, H. Menouar, and R. Hamila, "Revisiting crowd counting: State-of-the-art, trends, and future perspectives," *Image and Vision Computing*, vol. 129, p. 104597, Jan. 2023, doi:10.1016/j.imavis.2022.104597.

[3] D. Morgan, "Where are we?: camera movements and the problem of point of view," *New Review of Film and Television Studies*, vol. 14, no. 2, pp. 222–248, Feb. 2016, doi: 10.1080/17400309.2015.1125702.

[4] H. Rahmalan, M. S. Nixon, and J. N. Carter, "On crowd density estimation for surveillance," *IET Conference on Crime and Security*, vol. 2006, pp. 540–545, 2006, doi: 10.1049/ic:20060360.

[5] H. Rahmalan, N. Suryana, & N. A. Abu, "A general approach for measuring crowd movement," *Malaysian Technical Universities Conference and Exhibition on Engineering and Technology*, Jan. 2009.

[6] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim, "Recent survey on crowd density estimation and counting for visual surveillance," *Engineering Applications of Artificial Intelligence*, vol. 41, pp. 103–114, May 2015, doi: 10.1016/j.engappai.2015.01.007.

[7] Z. Fan, H. Zhang, Z. Zhang, G. Lu, Y. Zhang, and Y. Wang, "A survey of crowd counting and density estimation based on convolutional neural network," *Neurocomputing*, vol. 472, pp. 224–251, Feb. 2022, doi: 10.1016/j.neucom.2021.02.103.

[8] M. Elgendy, *Deep learning for vision systems*, Simon and Schuster, 2020.

[9] N. Sharma, R. Sharma, & N. Jindal, "Machine learning and deep learning applications-a vision," *Global Transitions Proceedings, 2(1)*, pp. 24-28, 2021.

[10] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, & J. Walsh, "Deep Learning vs. Traditional Computer Vision," *Advances in Computer Vision*, pp. 128–144, Apr. 2019, doi: 10.1007/978-3-030-17795-9_10.

[11] S. Islam et al., "A comprehensive survey on applications of transformers for deep learning tasks," *Expert Systems with Applications*, vol. 241, p. 122666, May 2024, doi:10.1016/j.eswa.2023.122666.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, & N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, & A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Advances in Neural Information Processing Systems, 34*, pp. 12116–12128, 2021.

[14] Y. Chen, J. Yang, B. Chen, and S. Du, "Counting Varying Density Crowds Through Density Guided Adaptive Selection CNN and Transformer Estimation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 3, pp. 1055–1068, Mar. 2023, doi:10.1109/tcsvt.2022.3208714.

[15] Y. Xiao et al., "A review of object detection based on deep learning," *Multimedia Tools and Applications*, vol. 79, no. 33–34, pp. 23729–23791, Jun. 2020, doi: 10.1007/s11042-020-08976-6.

[16] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, "Boosting Crowd Counting via Multifaceted Attention," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, doi:10.1109/cvpr52688.2022.01901.

[17] H. Lin, Z. Ma, X. Hong, Q. Shangguan, and D. Meng, "Gramformer: Learning Crowd Counting via Graph-Modulated Transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, pp. 3395–3403, Mar. 2024, doi: 10.1609/aaai.v38i4.28126.

[18] M. Zand, H. Damirchi, A. Farley, M. Molahasani, M. Greenspan, and A. Etemad, "Multiscale Crowd Counting and Localization By Multitask Point Supervision," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, doi: 10.1109/icassp43922.2022.9747776.

[19] C. Xu, K. Qiu, J. Fu, S. Bai, Y. Xu, and X. Bai, "Learn to Scale: Generating Multipolar Normalized Density Maps for Crowd Counting," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. 521, pp. 8381–8389, Oct. 2019, doi:10.1109/iccv.2019.00847.

[20] C. Liu, H. Lu, Z. Cao, and T. Liu, "Point-Query Quadtree for Crowd Counting, Localization, and More," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. 2105, pp. 1676–1685, Oct. 2023, doi: 10.1109/iccv51070.2023.00161.

[21] V. Sindagi, R. Yasarla, and V. M. M. Patel, "JHU-CROWD++: Large-Scale Crowd Counting Dataset and A Benchmark Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020, doi: 10.1109/tpami.2020.3035969.

[22] Q. Wang, J. Gao, W. Lin, and X. Li, "NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021, doi: 10.1109/tpami.2020.3013269.

[23] T. E. Oliphant, *Guide to NumPy (Vol. 1, p. 85)*, USA: Trelgol Publishing, 2006.

[24] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, Mar. 2011, doi:10.1109/mcse.2011.37.

[25] R. Shen, S. Bubeck, & S. Gunasekar, "Data augmentation as feature manipulation," *In International Conference on Machine Learning (PMLR)*, pp. 19773-19808, June, 2022.

[26] S.-H. Choi, "A study on Object Detection Method using Raspberry Pi," *Intelligent Information Convergence and Future Education*, pp.1-6, 2022.

[27] S.-H. Go, S.-M. Yang, H.-Y. Kim, and S.-B. Gwak, "Multi-Spectrum CNN-Based High-Resolution Color Image Interpolation Technique," *The Korean Association of Computer Education, 27(3)*, pp. 145-153, 2024.

[28] D. Kim, J. Jeon, S. Lim, and H. Lee, "An Object Pseudo-Label Generation Technique based on Self-Supervised Vision Transformer for Improving Dataset Quality," *Journal of KIISE*, vol. 51, no. 1, pp. 49–58, Jan. 2024, doi: 10.5626/jok.2024.51.1.49.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G., Chanan, & S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems, 32*, 2019.

[30] S. Imambi, K. B. Prakash, & G. R. Kanagachidambaresan, *PyTorch. Programming with TensorFlow: solution for edge computing applications*, pp. 87-104, 2021.

[31] Z. Ma, X. Hong, X. Wei, Y. Qiu, and Y. Gong, "Towards A Universal Model for Cross-Dataset Crowd Counting," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, doi:10.1109/iccv48922.2021.00319.