Enhancing Face Detection Performance in 360-Degree Video Using YOLOv8 with Equirectangular Augmentation Techniques

Rizky Damara Ardy [1)], Anny Yuniarti [2,*)], and Christy Atika Sari [3)]

[1, 2)] Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

[3)] Study Program in Informatics Engineering. Universitas Dian Nuswantoro, Semarang, Indonesia

E-mail: matthewvieri@yahoo.com[1)], anny@its.ac.id[2)], and atika.sari@dsn.dinus.ac.id[3)]

ABSTRACT

This study aims to enhance face detection performance in 360-degree videos by utilizing advanced image augmentation techniques with the YOLOv8 algorithm, which is effective for real-time object detection. Acknowledging the unique challenges posed by equirectangular projection, this research introduces a novel equirectangular augmentation method specifically designed for this medium. Our findings demonstrate a remarkable 1.346% improvement in detection accuracy in Equirectangular Projection (ERP) settings compared to standard YOLOv8 augmentation strategies. This significant enhancement not only addresses the geometric distortions inherent in panoramic video formats but also emphasizes the critical need for tailored augmentation approaches to improve face detection in complex environments. By showcasing the effectiveness of these customized methods, this research can contributes to the growing field of deep learning on face detection applications for immersive video technologies, with implications for sectors like security, navigation, and interactive. Ultimately, this work highlights the potential of innovative augmentation techniques to ensure robust face detection in challenging visual contexts.

Keywords: Face detection, 360-degree video, image augmentation, yolov8, deep learning, equirectangular projection, object detection, computer vision

1. Introduction

The rapid advancement in 360-degree video technology has significantly transformed how visual content is produced and consumed across various sectors, including entertainment, education, gaming, and security [1]. Unlike traditional video formats, which present a linear view of the world, 360-degree videos provide an immersive experience that allows viewers to look in any direction, resulting in a more engaging narrative [2]. Face detection in 360-degree videos is a role for supporting various activities. Navigation, a robot that used to map the presence of humans around it and navigate the surrounding environment safely [3]. Interaction, a 360-degree video users interact with virtual characters or other users detected in the video [4]. Analysis, [5]. As the popularity of this technology grows, so does the need for effective computer vision techniques that can operate within these unconventional visual environments.

Despite the potential applications, efficient and accurate face detection in immersive 360-degree videos presents unique challenges primarily due to the distortions inherent in this format. Traditional object detection algorithms, like those in the YOLO (You Only Look Once) family, often assume rectilinear and planar input. However, they struggle with the geometric distortions that arise from equirectangular projection, which transforms spherical representations of the environment into a rectangular format [6]. This transformation not only distorts the geometric properties of faces but also results in varying appearances depending on their position relative to the camera, often making faces skewed or disproportionately scaled [7]. Consequently, these challenges hinder YOLO models' capabilities to accurately identify and localize faces within 360 degree videos.

* Corresponding author.
Received: January 22nd, 2025. Revised: February 4th, 2025. Accepted: February 22nd, 2025.

Available online: February 25th, 2025.

DOI: 10.12962/j24068535.v23i1.a1255

In recent years, deep learning algorithms have emerged as powerful tools for various object detection tasks, including face detection. The YOLO (You Only Look Once) framework, particularly with its latest iteration, YOLOv8, has shown remarkable performance in real-time object detection scenarios [8]. YOLOv8 benefits from advancements in neural network architecture and training techniques, allowing it to achieve higher accuracy and faster inference speeds compared to its predecessors. This makes it an attractive option for dynamic environments such as 360-degree videos, where real-time processing is crucial. However, despite its impressive capabilities, the performance of YOLOv8 in the context of 360-degree face detection remains limited due to the aforementioned challenges.

This research seeks to enhance the robustness and performance of the YOLOv8 model by investigating the influence of tailored image augmentation techniques specifically catered to equirectangular formats. Image augmentation plays a vital role in training deep learning models, especially when the available datasets are limited or lack diversity [9]. By artificially expanding the training dataset, augmentation techniques help improve the model's generalization capabilities, enabling it to perform better in varied real-world conditions. This study aims to explore various augmentation strategies, including photometric transformations, geometric adjustments, and innovative blending techniques, to assess their impact on the face detection capabilities of YOLOv8 in equirectangular settings.

As the need for precise face detection in complex environments increases—whether for security applications, virtual reality experiences, or interactive media—it becomes imperative to develop and validate specialized methods that enhance detection accuracy while accounting for the unique characteristics of 360-degree videos. By understanding how different augmentation methods can be integrated into the training pipeline of YOLOv8, this research contributes to advancing deep learning applications for immersive video technologies and provides critical insights into optimizing face detection in challenging visual contexts.

Ultimately, this work emphasizes the potential of innovative augmentation techniques to ensure robust face detection, paving the way for more effective applications in various fields, including surveillance, augmented reality, and automated attendee tracking in events. The remaining sections of this paper detail the literature review, methodology, experimental results, and a conclusion that encapsulates our findings and implications for future research.

2. Related Works

The development of 360-degree video technology has significantly transformed the landscape of visual media, bringing forth new challenges in the field of computer vision. One of the principal concerns is the accurate detection of faces within these immersive environments, which has been deemed particularly challenging due to the geometric distortions introduced by equirectangular projections [7]. Research conducted by Yang et al. emphasized that traditional face detection methodologies are inadequate when applied to 360-degree video, leading to a call for innovative solutions that can manage the complexities of such visual data [2].

A promising approach lies in the application of deep learning algorithms for real-time face detection. Recent advancements, particularly regarding the YOLO framework, have enabled significant strides in object detection tasks due to its balance of speed and accuracy. YOLOv8, the latest version, has shown superior performance across a range of datasets. It is noted for its ability to effectively utilize context and spatial features to produce accurate detection results [8].

This study presents a comprehensive overview of face image augmentation, emphasizing the crucial role of dataset quality and size in deep learning tasks related to facial recognition. The challenges of collecting and labeling samples are addressed, alongside the widespread use of image augmentation techniques to enhance training datasets, as discussed in [9]. A systematic review of current literature on face image augmentation focuses on transformation types and methods, particularly deep learning approaches like Generative Adversarial Networks (GANs). Key transformations include geometric alterations that adjust image positioning and size, photometric adjustments for lighting and color enhancement, and hairstyle and makeup transfer techniques that allow for diverse appearances. Additional transformations, such as alterations in pose for varying viewpoints, expression synthesis for emotion

transfer, and age progression/regression for agebased variation, enrich the dataset. This research also explores challenges and opportunities within face image augmentation, such as identity preservation, control over image generation, and the diversity of produced images.

Image augmentation is a crucial technique in deep learning aimed at expanding image datasets by modifying existing images. This approach leads to better generalization, even with limited image collections [10]. Each augmentation technique generates new images that serve as additional variations for training data, thereby enhancing the performance of Convolutional Neural Networks (CNNs) [11]. The benefits of image augmentation include improving model accuracy by allowing it to learn a wider range of variations, thereby increasing its ability to recognize new situations. It also helps to reduce overfitting by providing more diverse images for the model to learn from, encouraging it to focus on general features rather than specific details. Furthermore, image augmentation boosts model robustness against noise by training the model to ignore irrelevant disturbances in the data [10].

There are various image augmentation techniques, each suited to different applications and types of image data. Common methods include photometric transformations, which adjust image color parameters such as brightness and contrast [12]; intensity transformations like noise addition and Random Erasing; geometric transformations that alter the image's positioning and shape; blending techniques that combine multiple images [12]; and pipeline augmentation that sequentially applies transformations for diverse training datasets [13]. These augmentation strategies are applicable in numerous deep learning tasks, such as object classification, detection, and segmentation, where they can enhance accuracy by generating images under varied conditions and clearer boundaries in medical applications [14]–[16].

The creation of deep learning models requires a dataset, which is vital for facilitating augmentation and face detection. Selecting an unsuitable dataset can negatively impact the effectiveness of face detection. A research, the authors employed the WiderFace dataset, a commonly used resource for face detection [17]. As such, this paper aims to build upon these foundational works and assess the performance enhancements garnered from applying advanced image augmentation methods in conjunction with YOLOv8 for face detection in 360-degree videos.

3. Methodology

The methodology for this study is structured into several key stages: dataset preparation, image augmentation, model training, and evaluation. Each stage plays a crucial role in ensuring a comprehensive enhancement of face detection performance in 360-degree videos utilizing the YOLOv8 algorithm. The following subsections elaborate on the processes involved in each stage, highlighting their importance and implementation.

3.1. Dataset Preparation

The dataset utilized in this research consists of images sourced from the WIDER Face dataset, a comprehensive resource well-regarded for its extensive annotations across a multitude of conditions, including variations in scale, pose, occlusion, and illumination [17]. This dataset is particularly valuable for training and evaluating face detection algorithms due to its diverse range of face appearances captured in various real-world scenarios. The WIDER Face dataset features over 32,203 images and includes more than 393,703 labeled faces, making it one of the largest datasets available for face detection tasks. These extensive annotations provide a robust foundation for training machine learning models, allowing them to learn the intricate characteristics associated with face detection across different environments and settings.

To ensure optimal compatibility with the YOLOv8 (You Only Look Once version 8) algorithm, the dataset underwent a rigorous refinement process tailored specifically for this study. WIDERFace annotations can be seen in Fig. 1, which have the following format:

- Line 1: Filename
- Line 2: Number of bounding boxes
- Line 3: Bounding box coordinates in the format of $x1$, $y1$, w, h, along with additional attributes such as blur, expression, illumination, invalid, occlusion, and pose. Here, $x1$ represents the x-coordinate of the bounding box, $y1$ represents the y-coordinate, w is the width, and h is the height of the bounding box. This is illustrated in Fig. 1.

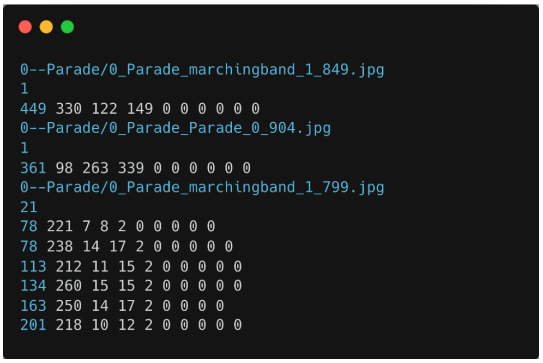


Fig. 1: Annotation sample WIDERFace

Because the annotations in the WIDERFace dataset differ in format from the YOLO format, the author performed preprocessing of the annotations using the following steps:

1. Load WIDERFace annotation data.
2. Lead the files; if a line containing ".jpg" is encountered, create a new file with the same name
3. Fill that file by searching for lines of data using regex and prepending a "0" to define the class of the face.

The YOLO annotation format is represented as follows:

```
<class_id> <x_center> <y_center> <width> <height>
```

For example, if we have a bounding box for a face with the coordinates ($x1$=0.48, $y1$=0.63) and dimensions ($width$=0.69, $height$=0.7) in normalized values, the corresponding entry in the YOLO format would look like:

```
0 0.48 0.63 0.69 0.7
```

In this representation, '0' indicates the class ID for the face, while 'x center', 'y center', 'width', and 'height' arenormalized by the dimensions of the image. This ensures that the YOLOv8 model receives the appropriate annotations for effective training.

The WIDER Face dataset is particularly valuable for training and evaluating face detection algorithms due to its diverse range of captured faces under various real-world scenarios. However, it is essential to acknowledge potential biases in the dataset that could affect the performance of YOLOv8 in 360-degree environments.

Potential biases in the dataset can arise from several factors, including:

- Demographic Bias: The WIDER Face dataset may not fully encompass the diversity of global populations. For instance, it might be overrepresented by certain ethnic groups or skin tones, which can lead to a model that performs well on faces closely resembling those in the dataset while struggling with diverse populations.
- Environmental Bias: The dataset is curated from various sources and conditions, but it may still demonstrate biases related to environmental factors, such as lighting and background. In a 360-degree video setting, where lighting conditions can change drastically, this could affect detection accuracy. For example, faces appearing in bright sunlight may be detected differently than those in dimly lit environments.
- Pose and Orientation Bias: The dataset may not adequately represent all possible angles and poses that a face can exhibit in a panoramic video. Given the nature of 360-degree environments, where faces can be viewed from any direction, limited representation of diverse orientations might hinder the YOLOv8 model's ability to detect faces accurately.

These biases can influence the model's generalization capabilities, potentially leading to decreased performance when deployed in real-world scenarios that differ from the training conditions. As a result, it is crucial to interpret the findings in light of these limitations, as they may affect the robustness of face detection in varied and dynamic environments. Future work could explore methods to augment training datasets further or utilize transfer learning techniques to alleviate these biases and enhance model robustness across diverse populations and conditions.

3.2. Image Augmentation

Image augmentation techniques were employed to significantly increase the diversity and richness of the training data, which is critical for enhancing the performance and generalization capabilities of the YOLOv8 model in detecting faces in 360-degree videos. The study implemented various augmentation methods, each designed to address specific challenges posed by the equirectangular format and the complexities associated with face detection. These methods included:

- Photometric Transformations: A range of adjustments in brightness, contrast, saturation, and color balance were systematically applied to simulate different lighting conditions present in real-world scenarios. By creating a dataset with varied illumination, the model gains robustness against the adverse effects of lighting changes, which are common in both indoor and outdoor environments. This form of augmentation is essential because it encourages the model to learn to identify and localize faces accurately, even when the lighting conditions are less than ideal, thereby enhancing its performance across different operational settings [9].
- Geometric Transformations: Techniques such as rotation, translation, scaling, and flipping were utilized to provide the model with different views and orientations of faces. These transformations are particularly beneficial in a 360-degree video context, where faces may appear from various angles due to the panoramic nature of the medium. By training on images that represent multiple orientations and scales, the model can better generalize its detection capabilities, ensuring high accuracy irrespective of how a face is oriented within the frame [18].
- Blending Augmentation: This approach involves blending two distinct images, along with their respective labels, to create new training samples. The implementation of blending augmentation includes selecting pairs of images randomly from the dataset and calculating a weighted sum of the pixel values to generate a composite image. By adjusting the blending ratio, we can control the degree of influence each image has on the final blended result. This technique not only enriches the dataset by merging features from multiple representations but also allows for the generation of diverse training samples that reflect various face appearances and contexts. This augmentation method is particularly beneficial for face detection in panoramic videos because it effectively simulates the complexities often encountered in real-world scenarios, such as occluded faces and partial visibility. In immersive environments, faces may appear from various angles, and blending allows the model to learn from a combination of features across different instances. Moreover, by exposing the model to a richer variety of facial attributes and backgrounds, blending augmentation assists in improving the generalization capabilities of the YOLOv8 model, ensuring robust face detection performance even in challenging visual contexts where panoramic distortions are present.
- Equirectangular Augmentation: Given the unique characteristics of 360-degree video, this technique creates images that closely resemble the equirectangular format. By simulating this projection style, the augmentation method prepares the model to better understand the spatial relationships and distortions inherent in 360-degree footage. This form of augmentation is particularly important because it addresses the geometric challenges that arise from panoramic captures, enabling the YOLOv8 model to effectively learn to navigate and detect faces within these complex environments. Unlike traditional augmentation techniques that typically apply linear transformations (e.g., rotations, translations) or photometric adjustments (e.g., brightness changes, color alterations) to rectilinear images, the equirectangular approach goes beyond these methods by accounting for the spherical geometry of the input data. This ensures that the augmentation maintains the proportionality and spatial integrity of visual features across varying viewpoints, ultimately leading to improved robustness in object detection tasks designed for immersive settings.

To achieve equirectangular augmentation, we utilize a series of mathematical transformations that convert spherical coordinates into Cartesian coordinates, and then map these to pixel values in the original image. The process begins by calculating longitude (ϕ) and latitude (θ) using the pixel indices (i, j) of the image. These are defined as:

$$\phi = \frac{j}{W_{eq}} \cdot 2\pi - \pi \tag{1}$$

Table 1: Summary Augmentation Technique

Category	Technique	Description
Photometric	Brightness	Changes the brightness level of images.
	HSV	Change Color Characteristic.
	BGR	Alters the intensity of colors in the image.
	Noise	Adds noise to create visual variability.
Geometric	Rotation	Rotates images to provide orientation variations.
	Translation	Moves images horizontally or vertically.
	Scaling	Changes the size of images.
	Shear	Shifting image horizontal or vertical.
	Flipping	Reverses images either horizontally or vertically.
	Perspective	Simulates different viewing angles.
	Mossaic	Collate more than one image into one image.
	Mixup	Merges features from two images to create training samples.
	Copy Paste	Overlay another image to an image
Equirectangular	Equirectangular	Adapts images for equirectangular projection to mitigate distortion.

$$\theta = \frac{i}{H_{eq}} \cdot \pi - \frac{\pi}{2} \tag{2}$$

Here, Weq and Heq designate the dimensions of the equirectangular image. Following this, we convert these spherical coordinates into Cartesian coordinates using:

$$x = cos(\theta) \cdot cos(\phi) \tag{3}$$

$$y = sin(\theta) \tag{4}$$

$$z = cos(\theta) \cdot sin(\phi) \tag{5}$$

These transformations allow for a comprehensive understanding of spatial arrangement, essential for training models to interpret 360-degree content effectively.

Once we have the Cartesian coordinates, the next step is to map them back to the appropriate pixel coordinates (u, v) in the original image. This step is crucial for ensuring that the augmented image accurately reflects the original scene. The mapping is accomplished using the following formulas:

$$u = clip\left(\left(\frac{x+1}{2}\right) \cdot W, 0, W-1\right) \tag{6}$$

$$v = clip\left(\left(\frac{y+1}{2}\right) \cdot H, 0, H-1\right) \tag{7}$$

Where W and H are the dimensions of the original image. This mapping not only helps maintain the fidelity of the original image during augmentation but also enables the model to effectively learn about the environmental context, thereby improving its capability to detect objects, such as faces, in complex 360-degree scenarios.

Summary of augmentation can show at Table 1, All of these augmentation techniques were seamlessly integrated into a comprehensive data processing pipeline using the Augmentor library in conjunction with the YOLOv8 framework. This integration not only facilitated a streamlined augmentation process but also ensured that a diverse and high-quality training dataset was produced, preserving the integrity of the original labels. Furthermore, for the equirectangular augmentation, researchers developed custom adjustments to incorporate specialized equirectangular code into the library, which was essential for simulating realistic 360-degree video conditions.

The YOLOv8 model was chosen for this study due to its commendable performance in real-time object detection applications [19]. The model was initialized with pre-trained weights sourced from the COCO dataset, which is widely recognized for its extensive coverage of object classes and intricate features. By leveraging these pre-trained weights, the model could capitalize on previously learned features, ultimately leading to faster convergence and improved detection accuracy during training.

YOLOv8 has capabilities other than detecting basic objects, namely offers diverse AI vision functialities, such as object detection, object classification, image segmentation, pose estimation, and object orientation [20]. YOLOv8 performance has increased, supported by an optimized backbone network, anchor-free detection head, and adjusted loss function [21]. YOLOv8 is a cutting-edge AI vision model that offers performance superior computing, high detection accuracy, and AI vision functionality diverse. With these advantages, YOLOv8 is an attractive solution for various applications in the fields of security, robotics, research and augmented reality [22].

Following the initialization, a tuning phase was conducted to optimize the hyperparameters using the ASHA (Asynchronous Successive Halving Algorithm) [23]. This tuning process aimed to identify the most effective settings that would enhance the model's performance while minimizing overfitting.

In terms of the parameters utilized for the augmentations, the research focused on refining configurations based on the default settings of YOLO. It is noteworthy that the equirectangular augmentation employed a simple binary toggle (on/off) system, which did not necessitate extensive parameter adjustments. This simplicity allowed for effective integration into the overall training workflow while ensuring that the benefits of the augmentation techniques were fully realized.

By employing these comprehensive image augmentation strategies, our methodology aimed to significantly enhance the learning capability of the YOLOv8 model, ultimately equipping it to perform robustly in the complex and dynamic context of 360-degree video environments.

3.3. Model Training

Following the successful optimization of hyperparameters through rigorous experimentation and analysis, the identified best parameters were employed to train the pre-trained model. This training phase was crucial in refining the model's ability to detect and classify objects effectively.

The training process began with the initialization of the model using pre-trained weights obtained from the COCO dataset. These weights provided a solid foundation, allowing the model to leverage previously learned features and patterns from a large and diverse dataset. By building upon this existing knowledge, the model could achieve faster convergence and improved performance on the specific task at hand.

During the training, the selected hyperparameters were meticulously applied, including augmentation parameters tailored to enrich the dataset. This dataset, enhanced through various augmentation techniques, was utilized to bolster the model's robustness by exposing it to diverse training scenarios, varying lighting conditions, and different face orientations. Such diversity in training data was vital for helping the model generalize effectively to unseen examples in real-world applications, which is critical for the deployment of face detection systems in dynamic environments.

Throughout the training phase, the model's performance was regularly monitored using several validation metrics, including precision, recall, and Mean Average Precision (mAP). The mAP, calculated as:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{8}$$

where N is the number of classes, AP_i and is the Average Precision for class i, served as a key indicator of the model's overall accuracy across various confidence thresholds [24]. This metric, along with others, ensured that the model was learning effectively. Adjustments were made as necessary based on the observed performance, guiding

the training process toward optimizing accuracy and reducing overfitting. The culmination of this comprehensive training and evaluation process led to a well-tuned model ready for deployment in real-time object detection tasks.

3.4. Evaluation

The effectiveness of the trained YOLOv8 model was evaluated using 360-degree video projections in Equirectangular Projection (ERP) format. This format provides a comprehensive view of the scene, allowing for an accurate assessment of the model's ability to detect faces in immersive environments. The evaluation process was structured around several key metrics that provide insights into the model's performance in detecting and localizing faces accurately.

The primary evaluation metrics used in this study included:

- Confidence: The confidence score refers to a probability value that the YOLOv8 model assigns to each predicted bounding box for an object detected in an image. This score indicates how certain the model is that the predicted bounding box contains a specific object [25], such as a face. Mathematically, the confidence score C for a predicted bounding box can be represented as:

$$C = P(Object \mid Image) \cdot I(O) \tag{9}$$

where $P(Object \mid Image)$ is the probability that an object exists in the bounding box according to the model, and $I(O)$ is the indicator function that is equal to 1 if the object class (for instance, "face") corresponds to the detected object, and 0 otherwise. A higher confidence score signifies a greater likelihood that the predicted bounding box accurately represents a face in the image.

- Count of Detected Objects: This metric involves counting the number of bounding boxes that appear in the testing video across various frames. The total detected face count N can be calculated as follows:

$$N = \sum_{i=1}^{k} T_i \tag{10}$$

where T_i represents the number of detected bounding boxes for image within the testing dataset, and k is the total number of test images or frames in the video. This metric is crucial, as it gives an overall quantification of how many faces the model was able to detect across the entire dataset.

Only confidence score and count of detected objects were selected as evaluation metrics due to the specific nature of our testing dataset, which primarily utilized an unlabeled collection of 360-degree videos. The reliance on unlabeled data limits the applicability of traditional performance metrics such as recall and F1-score, which require ground truth annotations to calculate true positives, false negatives, and false positives accurately.

By analyzing these metrics, the study aimed to identify the impact of the equirectangular image augmentation techniques on the detection performance of faces in 360-degree videos. The goal was to elucidate how well the YOLOv8 model could adapt to the unique challenges posed by equirectangular projections and assess the effectiveness of the implemented augmentation strategies in improving detection accuracy and robustness. This detailed evaluation phase provided insights into the model's strengths and limitations, guiding potential areas for further refinement and optimization in future research.

4. Experiment

4.1. Experiment Setup

This section details the experiments conducted to explain the process and evaluate the performance of equirectangular augmentation methods applied to YOLO (You Only Look Once) models. The primary objective was to assess the differences between default augmentation and default augmentation combined with equirectangular techniques. The experiments were conducted using three videos in Equirectangular Projection (ERP) format. Each video was evaluated based on metrics such as detection confidence and the count of detected objects.

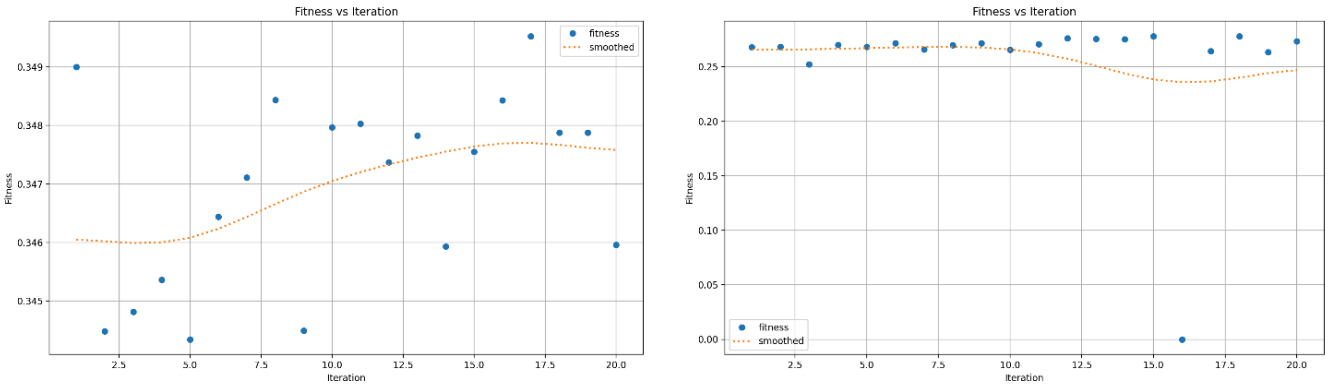
Table 2: Result Hyperparameter Standard and Equirectangular Augmentation

	Standard		**Equirectangullar**		**Standard**			**Equirectangular**		
					Fitness					
	Worst	**Best**	**Worst**	**Best**	**Min**	**Max**	**Range**	**Min**	**Max**	**Range**
iteration	5	17	3	15						
fitness	0.34434	0.34953	0.25224	0.27804	0.34434	0.34953	0.00519	0.25224	0.27804	0.0258
hsv_h	0.0134	0.01288	0.01398	0.02171	0.01288	0.0134	0.00052	0.01398	0.02254	0.00856
hsv_s	0.54387	0.63269	0.58999	0.50486	0.54387	0.63269	0.08882	0.50486	0.9	0.39514
hsv_v	0.33612	0.41169	0.47433	0.39402	0.33612	0.41169	0.07557	0.21858	0.52189	0.30331
degrees	0	0	0	0	0	0	0	0	0	0
translate	0	0	0	0	0	0	0	0	0	0
scale	0.11403	0.07974	0.10892	0.14067	0.07974	0.11403	0.03429	0.08323	0.14325	0.06002
shear	0.51634	0.4075	0.4545	0.42486	0.4075	0.51634	0.10884	0.32344	0.5	0.17656
perspective	0	0	0	0	0	0	0	0	0	0
flipud	0	0	0	0	0	0	0	0	0	0
fliplr	0.59409	0.46746	0.50933	0.40857	0.46746	0.59409	0.12663	0.17748	0.61417	0.43669
bgr	0	0	0	0	0	0	0	0	0	0
mosaic	0	1	1	0	0	1	0	0	1	0
mixup	0	0	0	0	0	0	0	0	0	0
copy_paste	0	0	0	0	0	0	0	0	0	0

4.2. Experiment Results

A. Hyperparameter

The process began with hyperparameter tuning of the default YOLO model alongside equirectangular augmentation. A total of 20 iterations, each consisting of 10 epochs, were conducted. The fitness score for the equirectangular augmentation showed a significant improvement, increasing from 0.25224 to 0.27804 over the iterations can show at Fig. 2. This indicates that the hyperparameter tuning effectively identified improved parameters. On the other hand, the default YOLO augmentation exhibited only a minor enhancement, indicated by a range of 0.00519 in fitness scores. Despite the slight increase, it still demonstrates some level of improvement.

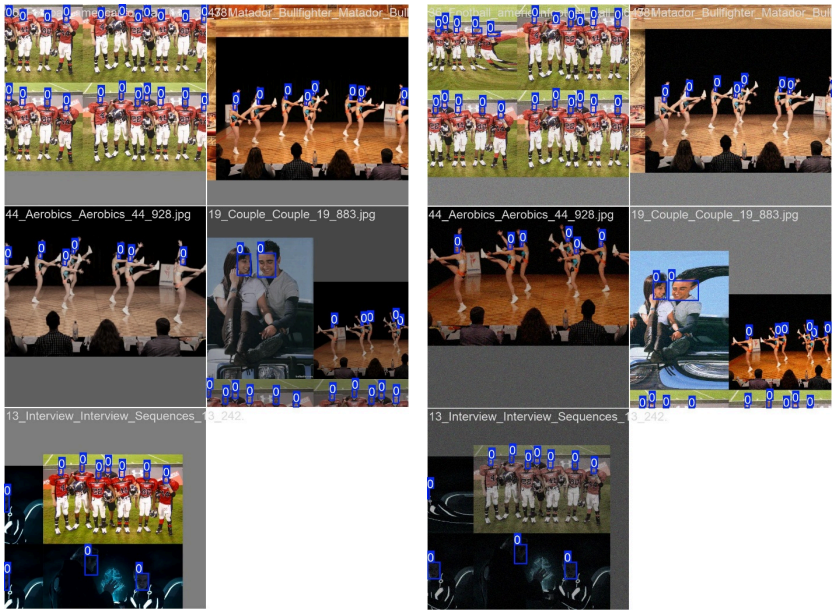


(a) Standard YOLO

(b) Equirectangular

Fig. 2: Tune Fitness Comparison

The results of the hyperparameter tuning for both the default YOLO and the equirectangular augmentation methods can be observed in Table 2. The best-performing parameters identified during this tuning process will be utilized for subsequent training.



(a) Standard YOLO (b) Equirectangular

Fig. 4: Performance Comparison Training

Table 3: Performance Comparison Training

No	Augmentation	mAP50	Best Epoch
1	Standard YOLO	0.68341	200
2	Equirectangular	0.71039	200

B. Training

Fig. 4 illustrates the differences in training samples between the standard YOLO and the Equirectangular augmentation methods. The performance results from these experiments are summarized in Table 3, which provides a comparative analysis of the training outcome.

From the table, we can observe distinct variances in mAP (mean Average Precision) scores between the two methods. The standard YOLO method achieved a mAP50 score of 0.68341 after 200 epochs, while the Equirectangular method demonstrated superior performance with a mAP score of 0.71039, also attained at 200 epochs. This indicates that the Equirectangular augmentation significantly enhances the model's ability to detect objects, resulting in a higher accuracy metric during validation.

Overall, the results suggest that incorporating Equirectangular augmentation into the training regimen allows for improved performance, particularly in scenarios where the model is required to process panoramic or spherical images. The comparative insights drawn from the data emphasize the potential benefits of choosing the appropriate augmentation techniques, as illustrated by the enhanced detection capabilities observed in the Equirectangular approach.

C. Testing Performance Model

The performance evaluation of augmentation methods in Equirectangular Projection (ERP) formats revealed insightful patterns in detection efficiency across various confidence ranges can show at Table 4 and the plot on Fig. 6. Notably, the initial ranges of 0.00 to 0.16 yielded no detections, because minium confidence set on 0.25.

The data becomes significant starting from the 0.24 - 0.2 range. Here, the Equirectangular augmentation outperformed Standard YOLO across various conditions. For instance, in ERP 1, the Equirectangular format detected 182 faces compared to 142 by Standard YOLO, demonstrating a more robust performance in face detection. Similarly, in ERP 2, Standard YOLO detected 1036 faces while Equirectangular detected 931. The results imply that

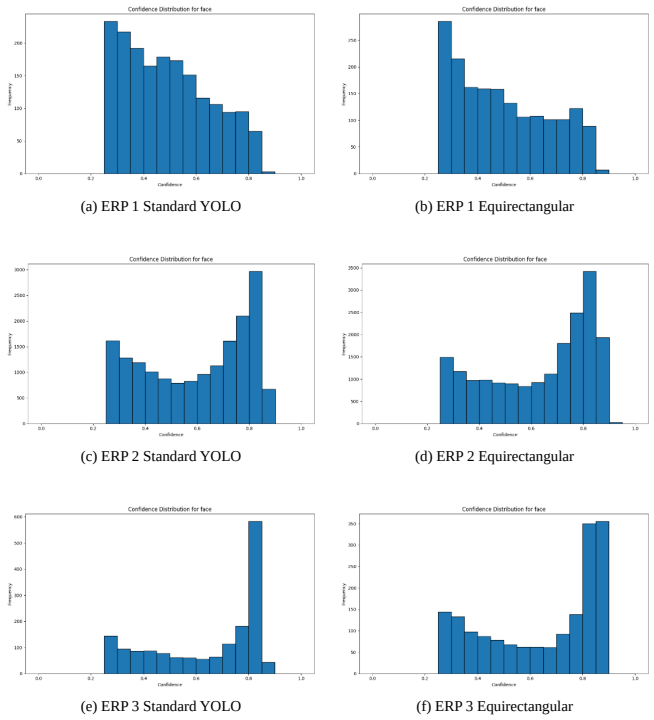
Table 4: Summary Confidence Performance

Range Conf.	Frequency						Total	
	ERP 1		ERP 2		ERP 3			
	Standard YOLO	Equirec.	Standard YOLO	Equirec.	Standard YOLO	Equirec.	Standard YOLO	Equirec.
0.00 - 0.04	0	0	0	0	0	0	0	0
0.04 - 0.08	0	0	0	0	0	0	0	0
0.08 - 0.12	0	0	0	0	0	0	0	0
0.12 - 0.16	0	0	0	0	0	0	0	0
0.16 - 0.20	0	0	0	0	0	0	0	0
0.20 - 0.24	0	0	0	0	0	0	0	0
0.24 - 0.28	142	182	1036	931	98	98	1276	1211
0.28 - 0.32	182	184	1092	1071	88	104	1362	1359
0.32 - 0.36	169	173	1020	852	67	97	1256	1122
0.36 - 0.40	149	123	938	779	71	75	1158	977
0.40 - 0.44	122	132	847	782	62	71	1031	985
0.44 - 0.48	150	118	712	742	73	60	935	920
0.48 - 0.52	148	123	652	729	49	61	849	913
0.52 - 0.56	130	99	652	705	47	54	829	858
0.56 - 0.60	118	83	633	668	55	49	806	800
0.60 - 0.64	100	83	761	719	44	51	905	853
0.64 - 0.68	83	94	867	830	46	48	996	972
0.68 - 0.72	86	68	1095	1092	76	53	1257	1213
0.72 - 0.76	58	88	1353	1698	93	90	1504	1876
0.76 - 0.80	84	99	1718	1995	156	111	1958	2205
0.80 - 0.84	53	78	2432	2531	499	221	2984	2830
0.84 - 0.88	15	18	1196	2376	125	463	1336	2857
0.88 - 0.92	0	0	10	467	1	21	11	488
0.92 - 0.96	0	0	0	0	0	0	0	0
0.96 - 1.00	0	0	0	0	0	0	0	0
Average	0.4923	**0.4949**	0.6001	**0.6356**	0.6437	**0.6459**	0.5787	**0.5922**
Face Count	1789	**1745**	17014	**18967**	1650	**1727**	20453	**22439**

while Standard YOLO was effective, the Equirectangular approach provided a slight edge in detection frequency, especially at this mid-range confidence level.

Moreover, the data highlights a positive correlation between confidence level and detection counts. For example, in subsequent ranges, ERP 2 showed average performance values climbing from 0.6001 (Standard YOLO) to 0.6356 (Equirectangular), reinforcing the argument that Equirectangular augmentation enabled higher confidence detections.

The results indicate a total detected face count across all formats reached 20,453, with the Equirectangular augmentation leading to 22,439 detections, which represents a 10% increase in the total count of detected faces across various 360-degree video projections. This improvement is particularly significant given the inherent complexities associated with face detection in immersive environments. Traditional object detection algorithms often struggle with the distortions presented by equirectangular projections. By enhancing detection capabilities, this study not only demonstrates the effectiveness of tailored augmentation strategies but also ensures greater reliability



(a) ERP 1 Standard YOLO (b) ERP 1 Equirectangular

(c) ERP 2 Standard YOLO (d) ERP 2 Equirectangular

(e) ERP 3 Standard YOLO (f) ERP 3 Equirectangular

Fig. 6: Performance Comparison Training

in real-world applications, such as surveillance, human-computer interaction in virtual reality, and social behavior analysis in crowded public spaces. The ability to detect more faces more accurately can lead to better situational awareness and responsiveness in various technological deployments, underscoring the importance of advancing methodologies that adapt to challenges specific to 360-degree imaging.

From the comparative analysis of specific video clips on Fig. 7, the inference supported the data-driven observations. In ERP 1, Standard YOLO detected no faces, while Equirectangular detected 3 faces, demonstrating a clear superiority of the latter in challenging conditions. Throughout ERP 2, both methods showed equality in face detection, revealing situations where both augmentations were equally effective. In ERP 3, the consistent detection of 2 faces by both methods illustrates their comparable capabilities when confronted with certain video complexities.

These results emphasize the significant impact of data augmentation strategies on the performance of YOLO models in detecting objects in different video projections. Future work can explore further optimal configurations for augmentations to enhance model performance and effectiveness.

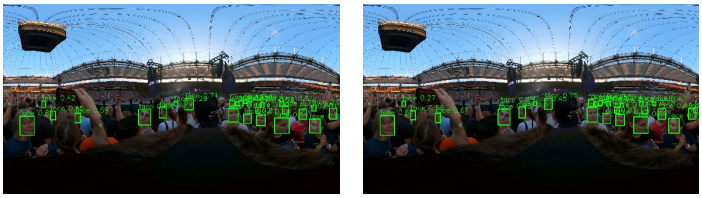
5. Conclusion

This study highlighted the critical importance of image augmentation in improving the performance of face detection within 360-degree videos using the YOLOv8 algorithm. The research demonstrated that equirectangular augmentation techniques notably enhance the model's performance. The results revealed an impressive improvement in detection performance, with the equirectangular technique leading to increases of 1.346% for Equirectangular Projection (ERP) when compared to default augmentation practices employed by YOLOv8.

The findings of this study indicate that carefully designed augmentation strategies can effectively address the unique challenges posed by 360-degree video environments, characterized by significant visual distortions. Future work could further explore additional augmentation techniques and their combinations to refine face detection capabilities, ensuring the model's robustness across diverse real-world scenarios.

In conclusion, the successful incorporation of image augmentation techniques, particularly equirectangular, has broad implications for enhancing automated face detection systems in immersive video formats, paving the way for safer and more interactive applications in fields such as security and virtual reality.

(a) ERP 1 Standard YOLO (b) ERP 1 Equirectangular



(c) ERP 2 Standard YOLO (d) ERP 2 Equirectangular

(e) ERP 3 Standard YOLO (f) ERP 3 Equirectangular

Fig. 7: Video Inference Comparison

CRediT Authorship Contribution Statement

Rizky D. Ardy: Conceptualization, Methodology, Software, Investigation, Resources, Data Curation, Writing – Original Draft, Visualization, Funding Acquisition. **Anny Yuniarti:** Validation, Formal analysis, Resources, Writing – Review & Editing, Supervision, Project Administration. **Christy A. Sari:** Writing – Review & Editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The dataset was openly provided [http://shuoyang1213.me/WIDERFACE/index.html].

Declaration of Generative AI and AI-assisted Technologies in The Writing Process

The authors used generative AI to improve the writing clarity of this paper. They reviewed and edited the AI-assisted content and take full responsibility for the final publication.

References

[1] M. Xu, C. Li, S. Zhang, and P. L. Callet, “State-of-the-art in 360° video/image processing: Perception, assessment and compression,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 1, pp. 5–26, Jan. 2020, ISSN: 1941-0484. DOI: 10.1109/JSTSP.2020.2966864.

[2] C.-Y. Yang and H. H. Chen, “Efficient face detection in the fisheye image domain,” IEEE Transactions on Image Processing, vol. 30, pp. 5641–5651, 2021. DOI:10.1109/TIP.2021.3087400.

[3] A. Bacchin, F. Berno, E. Menegatti, and A. Pretto, “People tracking in panoramic video for guiding robots,” in Lecture Notes in Networks and Systems. Springer Nature Switzerland, 2023, pp. 407–424, ISBN: 9783031222160. DOI: 10.1007/978-3-031-22216-0 28. [Online].

[4] X. Pan and A. F. d. C. Hamilton, “Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape,” *British Journal of Psychology*, vol. 109, no. 3, pp. 395–417, Mar. 2018, doi: 10.1111/bjop.12290.

[5] A. Doula, A. Sanchez Guinea, and M. M¨uhlh¨auser, “Vr-surv: A vr-based privacy preserving surveillance system,” in Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, ser. CHI EA ’22, New Orleans, LA, USA: Association for Computing Machinery, 2022, ISBN: 9781450391566. DOI:10.1145/3491101.3519645. [Online].

[6] J. Fu, S. Ranjbar Alvar, I. Bajic, and R. Vaughan,“Fddb-360: Face detection in 360-degree fisheye images,” in 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 15–19. DOI: 10.1109/MIPR.2019.00011.

[7] R. G. d. A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli, and P. Frossard, “Visual distortions in 360° videos,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 8, pp. 2524–2537, 2020. DOI: 10 . 1109 / TCSVT . 2019 .2927344.

[8] W. Yang and Z. Jiachun, “Real-time face detection based on yolo,” in 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), 2018, pp. 221–224. DOI: 10.1109/ICKII.2018.8569109.

[9] X. Wang, K. Wang, and S. Lian, “A survey on face data augmentation for the training of deep neural networks,”Neural Computing and Applications, vol. 32, no. 19, pp. 15 503–15 531, Mar. 2020, ISSN: 1433-3058. DOI:10 . 1007 / s00521 - 020 - 04748 - 3. [Online].

[10] J.-H. Lee, M. Z. Zaheer, M. Astrid, and S.-I. Lee, “SmoothMix: a Simple Yet Effective Data Augmentation to Train Robust Classifiers,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2020, pp. 3264–3274, doi: 10.1109/cvprw50498.2020.00386.

[11] M. Kumar and U. Mehta, “Enhancing the performance of CNN models for pneumonia and skin cancer detection using novel fractional activation function,” *Applied Soft Computing*, vol. 168, p. 112500, 2025, doi: 10.1016/j.asoc.2024.112500.

[12] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, “A Comprehensive Survey of Image Augmentation Techniques for Deep Learning,” Pattern Recognition, vol. 137, p. 109347, May 2023, doi: 10.1016/j.patcog.2023.109347.

[13] N. Groun, M. Villalba-Orero, L. Casado-Martín, E. Lara-Pezzi, E. Valero, J. Garicano-Mena, and S. Le Clainche, “A novel data augmentation tool for enhancing machine learning classification: A new application of the higher order dynamic mode decomposition for improved cardiac disease identification,” *Results in Engineering*, vol. 25, p. 104143, 2025, doi: 10.1016/j.rineng.2025.104143.

[14] Y. Liu, X. Wang, Z. Zhang, and F. Deng, “Deep learning based data augmentation for large-scale mineral image recognition and classification,” *Minerals Engineering*, vol. 204, p. 108411, 2023, doi: 10.1016/j.mineng.2023.108411.

[15] H. Jo, Y.-H. Na, and J.-B. Song, “Data augmentation using synthesized images for object detection,” in 2017 17th International Conference on Control, Automation and Systems (ICCAS), 2017, pp. 1035–1038. DOI: 10.23919/ICCAS.2017.8204369.

[16] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, “Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 8535–8545, doi: 10.1109/cvpr.2019.00874.

[17] S. Yang, P. Luo, C. C. Loy, and X. Tang, “WIDER FACE: A Face Detection Benchmark,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 0, no. , pp. 5525–5533, doi: 10.1109/CVPR.2016.596.

[18] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” Journal of Big Data, vol. 6, no. 1, p. 60, 2019, ISSN: 2196-1115. DOI:10.1186/s40537-019-0197-0. [Online].

[19] M. Sohan, T. Sai Ram, and C. V. Rami Reddy, “A Review on YOLOv8 and Its Advancements,” in *Data Intelligence and Cognitive Informatics, Springer Nature Singapore*, 2024, pp. 529–545.

[20] A. Rahim, F. Yuan, and J. Barabady, “An Ultralytics YOLOv8-Based Approach for Road Detection in Snowy Environments in the Arctic Region of Norway,” *Computers, Materials and Continua*, vol. 83, no. 3, pp. 4411–4428, 2025, doi: 10.32604/cmc.2025.061575.

[21] A. Batool, Y.-W. Kim, and Y.-C. Byun, “Improved YOLOv8 framework for efficient solar panel defect detection,” *Journal of Building Engineering*, vol. 111, p. 113031, 2025, doi: 10.1016/j.jobe.2025.113031.

[22] S. Li, H. Schieber, N. Corell, B. Egger, J. Kreimeier, and D. Roth, “GBOT: Graph-Based 3D Object Tracking for Augmented Reality-Assisted Assembly Guidance,” in *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 2024, pp. 513–523, doi: 10.1109/VR58804.2024.00072.

[23] L. Li et al., “A System for Massively Parallel Hyperparameter Tuning,” in *Proceedings of Machine Learning and Systems*, 2020, vol. 2, pp. 230–246, [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2020/file/a06f20b349c6cf09a6b171c71b88bbfc-Paper.pdf.

[24] Y. C. Keluskar, N. G. Singhaniya, V. A. Vyawahare, C. S. Jage, P. Patil, and G. Espinosa-Paredes, “Solution of nonlinear fractional-order models of nuclear reactor with parallel computing: Implementation on GPU platform,” *Annals of Nuclear Energy*, vol. 195, p. 110134, 2024, doi: 10.1016/j.anucene.2023.110134.

[25] J. Du, “Understanding of object detection based on cnn family and yolo,” Journal of Physics: Conference Series, vol. 1004, no. 1, p. 012 029, Apr. 2018. DOI: 10.1088/1742-6596/1004/1/012029. [Online].