Explainable Object Detection in 360° Images Through Fuzzy Logic Systems and Vision-Language Models Integration

Amer Harfoush, Hani Hagras
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
ah24491@essex.ac.uk

Hugo Leon-Garza, Anasol Pena-Rios
Business Modelling and Operational Transformation Practice
British Telecom
Ipswich, UK
anasol.penarios@bt.com

***Abstract*—Panoramic images (360-degree) contain significant amount of information since they manage to take a snapshot of the complete surroundings at once which makes them a great choice in multiple fields. However, they pose special challenges for object detection due to significant geometric distortions coming from equirectangular projection. Current methods are inadequate for critical applications since they do not have the needed accuracy and provide no explanations for their decisions. This paper presents a novel framework that integrates Fuzzy Logic Systems (FLSs), an object detection model, and Vision-Language Models (VLMs) to achieve both good detection performance and explainability. Our method uses a fine-tuned YOLOv9 object detector enhanced by a VLM-based validation engine. The fuzzy logic system module evaluates image quality and adapts processing parameters, providing linguistic explanations that bridge numerical processing with human understanding. Testing on industrial datasets from British Telecom demonstrates robust performance on previously unseen object categories, validating real-world applicability. The proposed framework achieves up to 124.67% improvement over previous state-of-the-art methods while providing explanations for each detection decision.**

***Keywords*—** *fuzzy logic, explainable AI, panoramic vision*

I. Introduction

The increasing availability and affordability of 360-degree cameras has revolutionized how we capture and analyse environments, offering complete omnidirectional views from a single vantage point. These devices have found applications in autonomous robotics, virtual reality, facility management, and surveillance [1]. However, the equirectangular projection (ERP) format used to represent spherical data on flat images introduces severe geometric distortions that fundamentally challenge traditional computer vision approaches. Objects near the poles appear stretched horizontally by factors up to 10×, while those near the equator maintain relatively normal proportions [2]. This non-linear distortion means a chair might appear normal in one image region but completely warped in another. Standard object detectors trained on perspective images fail when confronted with panoramic distortions, often missing objects entirely or generating numerous false positives [3], [4]. More critically, existing approaches even those specifically designed for 360° imagery operate as black boxes, providing no explanation for their decisions. When a detector claims there is a "chair" with 73% confidence, we have no understanding of why it believes that or whether distortion affected its judgment. State-of-the-art methods achieve only 22.7% mAP@50 on the PANDORA benchmark [3], demonstrating the magnitude of the technical challenge.

The lack of explainability becomes particularly problematic in industrial applications. Consider a facility management system inventorying equipment across multiple sites. If the system misses critical safety equipment or falsely identifies objects, understanding why these errors occurred is essential for system improvement and user trust. Current approaches offer no such insights, limiting their deployment in safety-critical domains where accountability and transparency are crucial [4].

Previous research investigated several approaches to 360° object detection including Projection-based methods which convert equirectangular images to perspective views (cube faces, tangent planes) before applying standard detectors [5], [6]. While this reduces distortion, it fragments the scene, causes objects crossing projection boundaries to be detected multiple times, and requires expensive multi-view processing without sharing intermediate representations [7]. On the other hand, Distortion-aware architectures modify convolutional neural networks to account for spherical geometry; for example SphereNet [8] adapts sampling locations of convolutional filters to encode distortion invariance, while spherical CNNs [9] define operations directly on spherical harmonics. These methods show improvements but face high computational complexity and difficulty leveraging powerful pre-trained models due to architectural incompatibilities.

Specialized representations introduce novel annotation formats like Bounding Field-of-View (BFoV) [10] and Rotated BFoV (RBFoV) [3] that better fit distorted objects. R-CenterNet, current state-of-the-art, uses oriented heatmap generation and panoramic rotation augmentation [3]. However, the performance remains far from planar detection benchmarks (50-60% mAP on COCO), and the system provides no explanations for its predictions.

All existing methods share critical limitations in that they treat 360° detection purely as a geometric problem without addressing the fundamental uncertainty in distorted feature interpretation. Furthermore, they provide only numerical confidence scores without semantic justification, and they cannot handle the multi-layered uncertainty arising from sensor noise, varying illumination, and ambiguous boundaries in highly distorted regions.

This paper addresses both the technical challenge of distortion and the practical need for explainability through a novel integration of three complementary technologies: fine-tuned deep learning model for feature extraction, fuzzy logic for uncertainty management, and a vision-language model for semantic validation. The proposed approach achieves 48.4% mAP@50 on PANDORA dataset a 113% improvement over previous state-of-the-art (22.7%) and 51.04% mAP@50 on 360-Indoor dataset, representing a 124.67% improvement over the baseline (22.8%). We demonstrate robust generalization on British Telecom's industrial dataset containing object categories never seen during training, proving the framework's real-world applicability for facility management and autonomous inspection.

The paper is organized as follows. Section II provides brief background on 360° panoramic images and the challenges they pose. Section III presents the proposed system integrating fuzzy logic with object detection and VLM validation. Section IV presents comprehensive experiments and results. Section V concludes with future research directions.

II. Brief Overview on 360° Panoramic Imaging

A. Equirectangular Projection and Geometric Distortions

360-degree images capture the complete spherical field of view surrounding a camera through omnidirectional lenses. As shown in Fig. 1, the raw version of the image captured is very hard to be used by any means.

Fig. 1. Spherical image capture by 360° camera [7].

The most common representation is the equirectangular projection (ERP), which maps spherical coordinates (θ, φ) to Cartesian image coordinates (x, y) through [11]:

$$x = \frac{W}{2\pi}(\theta + \pi) \quad (1)$$

$$y = \frac{H}{\pi}(\frac{\pi}{2} - \phi) \quad (2)$$

where W and H are image width and height, $\theta \in [-\pi, \pi]$ is longitude (azimuth angle), and $\varphi \in [-\pi/2, \pi/2]$ is latitude (elevation angle) [3]. Fig. 2 shows the ERP and how it is hard to deal with the image in that state right away. This mapping introduces severe non-linear distortions that vary continuously with latitude. The distortion factor at latitude φ is proportional to $1/\cos(\varphi)$, causing objects to appear increasingly stretched as they approach the poles ($\varphi \rightarrow \pm 90^\circ$) [11].

Fig. 2. Equirectangular projection of 360° camera [7].

Mathematically, a spherical object of radius r centred at latitude φ appears in the equirectangular plane with dimensions:

$$w_{apparent} = \frac{2r}{\cos(\emptyset)} \quad (3)$$

$$h_{apparent} = 2r \quad (4)$$

where $W\_apparent$ grows unbounded as $\phi \rightarrow \pm 90^\circ$. This means identical objects appear radically different depending on their position where a circular ceiling light near the pole becomes an extreme horizontal ellipse, while the same light near the equator maintains near-circular appearance.

B. Challenges for Object Detection

The continuous distortion gradient creates three fundamental problems for computer vision. First, standard convolutional neural networks learn translation-invariant features, assuming objects maintain consistent appearance regardless of position. This assumption fails in panoramic images where the same object class exhibits massively different aspect ratios depending on latitude. A chair at the equator might have aspect ratio 1:1, but at 80° latitude, distortion stretches it to 6:1, appearing as an unrecognizable horizontal band.

Second, image boundaries represent a topological challenge. The leftmost and rightmost columns are the same point ($\theta = \pm 180°$) in spherical space, meaning objects crossing this boundary appear split into two disconnected regions in the planar representation [10]. Standard detectors trained on perspective images have never encountered this "wrap-around" topology and consistently fail to recognize such split objects, either missing them entirely or detecting each fragment as a separate instance.

C. Previous Distortion Handling Approaches

Prior work attempted strategies to mitigate these challenges. Projection-based methods convert the equirectangular image to N perspective views (commonly N=6 cube faces or N=12-20 tangent planes) before applying standard detectors [5], [6]. While each perspective view has minimal distortion, this approach requires N forward passes through the detector without sharing intermediate features, increasing computational cost N-fold. More critically, objects appearing in multiple views generate duplicate detections and objects centred on projection seams may be split with incomplete appearance [7].

Architecture modification methods design convolutional kernels adapted to spherical geometry. Distortion-aware CNNs [11] sample non-regular grids based on each pixel's distortion level while maintaining parameter sharing, and spherical CNNs

[9] define operations directly in spherical harmonic space to avoid distortions entirely. However, these approaches face limitations: they cannot leverage powerful pre-trained models due to incompatible architectures, forcing training from scratch with limited panoramic datasets [9].

Our approach differs fundamentally: rather than attempting to eliminate distortions through projection or specialized architectures, we train directly on equirectangular images and use fuzzy logic to model the continuous uncertainty gradient caused by varying distortion levels.

III. The Proposed Fuzzy Logic System and Vision-Language Models Integration for 360° Images

A. System Architecture

Our framework (shown in Fig. 3) consists of three primary components working together. The base detector processes the full equirectangular image in a single forward pass [12], generating initial detections with associated confidence scores. The fuzzy logic system evaluates in parallel the image characteristics, producing linguistic assessments, e.g., "*poor quality due to low contrast and high distortion*". These assessments, along with the initial detections, feed into the VLM which validates each detection, suggests missing objects, and provides detailed explanations for all decisions taking into consideration the spatial relationships between objects to verify if the object is in the correct place or is contextually implausible for example: "*A car in a living room*". Fig. 3 shows the integration of fine-tuned detection, fuzzy logic quality assessment, and VLM validation.

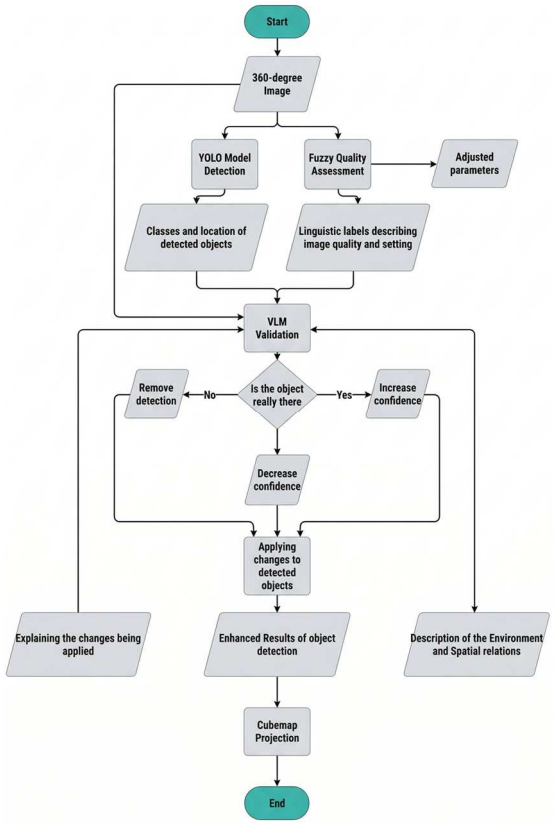

Fig. 3. Overall architecture of the proposed system.

B. Fuzzy Logic Quality Assessment

The employed fuzzy system has five inputs (Brightness, Contrast, Sharpness, Distortion Level, Colour Saturation) and three outputs (Quality, Difficulty, Adjustment). Each input employs multiple fuzzy sets. Fig. 4 shows an example of the employed membership functions for quality.

Fig. 4. Quality triangular membership functions.

We designed 15 rules through analysis of failed detections on validation data. The employed rule base is shown in Table I.

TABLE I. Rules from our Fuzzy Logic Rule Base System

	INPUT					OUTPUT		
Rule	Brightness	Contrast	Sharpness	Distortion	Saturation	Quality	Difficulty	Adjustment
1	VERY_DARK	LOW	-	-	-	POOR	HARD	LOW
2	NORMAL	HIGH	SHARP	-	-	EXCELLENT	EASY	HIGH
3	-	-	MODERATE	HIGH	-	MODERATE	HARD	MODERATE
4	-	LOW	-	-	LOW	POOR	HARD	LOW
5	DARK	-	SHARP	HIGH	HIGH	MODERATE	HARD	LOW
6	BRIGHT	-	-	HIGH	-	MODERATE	HARD	MODERATE
7	NORMAL	MEDIUM	SHARP	HIGH	-	GOOD	MODERATE	MODERATE
8	NORMAL	MEDIUM	SHARP	LOW	HIGH	EXCELLENT	EASY	HIGH
9	NORMAL	MEDIUM	MODERATE	MEDIUM	MEDIUM	MODERATE	MODERATE	MODERATE
10	-	-	BLURRY	LOW	-	POOR	HARD	LOW
11	NORMAL	HIGH	MODERATE	LOW	HIGH	GOOD	EASY	HIGH
12	DARK	HIGH	SHARP	LOW	-	MODERATE	MODERATE	MODERATE
13	BRIGHT	LOW	-	-	-	POOR	HARD	LOW
14	NORMAL	MEDIUM	SHARP	LOW	LOW	GOOD	EASY	HIGH
15	DARK	MEDIUM	BLURRY	HIGH	LOW	POOR	HARD	LOW

Determining optimal membership function parameters required combining domain expertise with empirical validation. The initial membership function parameters were established through analysis of training dataset characteristics. For brightness assessment, we examined histogram distributions of 500 randomly sampled images, identifying natural clusters corresponding to very dark (mean intensity <60), dark (60-130), normal (110-200), and bright (>180) categories. Triangular membership functions were positioned with peaks at these cluster centres and overlaps designed to capture gradual transitions. Similarly, contrast thresholds were derived from standard deviation distributions, sharpness levels from Laplacian variance analysis, and distortion zones from the geometric properties of equirectangular projection.

The 15 fuzzy rules were crafted through systematic analysis of detection failures on the validation set. We manually inspected ~200 failed detections, categorizing failure modes by their associated image characteristics. For instance, images with brightness <70 and contrast <35 consistently produced unreliable detections, motivating Rule 1: "IF brightness is VERY_DARK AND contrast is LOW THEN quality is POOR." Each rule encodes a specific failure pattern identified through this qualitative analysis.

After implementing the initial fuzzy logic system, we conducted iterative refinement through an empirical process. The process involved running detection on a held-out validation subset of 300 images, manually reviewing results to identify systematic errors, adjusting parameters, and re-running detection to measure mAP@50 improvement. For example, we initially set the MEDIUM contrast range as μ (x; 20, 45, 70), but observed misclassification of high-contrast images. Shifting to μ (x; 25, 50, 75) improved detection by 3.2% mAP. The parameters converged to stable values producing 47.3% mAP@50, a 7.3% improvement over the baseline (40.0%). Validation on the held-out test set achieved 48.4% mAP@50 with VLM integration, confirming generalization.

C. VLM Integration and Validation

The engineered prompt incorporates:

1. Current detections with spherical coordinates
2. Fuzzy logic quality assessment
3. Expected distortion patterns for 360° imagery
4. Task-specific validation criteria

D. Fine-Tuned Detection Model

Standard YOLO training on perspective images fails for panoramic data due to the geometric distortions discussed in Section II. Our fine-tuning approach makes three critical adaptations. First, we convert RBFoV annotations ($\theta, \varphi, \alpha, \beta, \gamma$) used in panoramic datasets to YOLO's expected format (x_{centre}, y_{centre}, *width, height*) while accounting for latitude-dependent stretching:

$$x_{centre\_norm} = \frac{\theta+\pi}{2\pi} \tag{5}$$

$$y_{centre\_norm} = \frac{\pi/_2-\emptyset}{\pi} \tag{6}$$

$$w_{norm} = \frac{\propto}{360} \times \frac{1}{\cos(\emptyset)} \tag{7}$$

$$h_{norm} = \frac{\beta}{180} \tag{8}$$

where θ is the longitude of the object centre in radians, $\emptyset$ is the latitude of the object center in radians, where α and β are the horizontal and vertical fields of view in degrees. x_{centre_norm} and y_{centre_norm} are the normalized centre coordinates in image space, and w_{norm} is the normalized bounding box width after distortion compensation. The crucial term $\frac{1}{\cos(\emptyset)}$ compensates for latitude-dependent horizontal compression, ensuring the model learns actual object extents rather than projection artifacts [11].

For the inverse transformation from image space to spherical coordinates the following two equations are used:

$$\theta = \frac{2\pi x}{W} - \pi \tag{9}$$

$$\emptyset = \frac{\pi}{2} - \frac{\pi y}{H} \tag{10}$$

Where W and H are the image's width and Height respectively in pixels. This transformation accounts for latitude-dependent stretching, ensuring the model learns actual object patterns rather than projection artifacts. Training for 30 epochs using horizontal flip, mosaic, mix-up, and HSV colour jittering, with carefully disabled augmentations that would violate spherical geometry (rotation, perspective transforms).

E. Fuzzy Guided Adaptive Processing

The linguistic outputs from the FLS directly modify detection parameters as follows:

IF quality IS poor THEN threshold = threshold_base × 0.85
IF difficulty IS hard THEN nms_threshold = nms_base × 1.15
IF adjustment IS high THEN search_region = region_base × 1.20

This adaptive approach allows the system to handle varying conditions without manual intervention. This approach provides the necessary information for the user to walk through the process and provide all the linguistic labels for each image to the VLM to deal with each image depending on its settings.

F. Vision-Language Model Integration

The VLM component represents a significant advancement. Rather than using it for primary detection, we employ Gemini-1.5-Flash as a validation and explanation engine. It receives the fuzzy linguistic assessment along with detection results and the ERP image for example:

" Image Quality: MODERATE (μ=0.65)
Brightness: POOR (darkness=85, high shadows)
Distortion: HIGH (polar regions affected)
Recommendation: Apply enhanced processing for φ > 75° "

This linguistic context enables the VLM to understand detection conditions and provide appropriate validation. For each detection, the VLM provides validation status with reasoning, confidence adjustment *(0.5-1.5x),* and specific explanation for the decision. When it identifies missing objects with high confidence, the system attempts re-detection with lowered thresholds in suggested areas. The explanations aren't generic, they're specific and actionable. Instead of just "removed false positive," the system reports "removed duplicate sink detection caused by image boundary split at *θ=180°*". This level of detail transforms debugging from guesswork to informed refinement. The VLM uses spatial relationships between detections and each other to find common sense between these objects, helping it recognize false detections more often.

To be able to connect objects, we use angular distances between objects:

$$angular\_distance = arccos(sin(\varphi_1) \times sin(\varphi_2) + cos(\varphi_1) \times cos(\varphi_2) \times cos(\theta_1 - \theta_2)) \tag{11}$$

The VLM can explain decisions such as "Chair confidence boosted despite poor lighting, as four-leg structure clearly visible." where (φ_1, θ_1) and (φ_2, θ_2) are the latitude and longitude coordinates of two detected objects in radians.

IV. Experiments and Results

A. Dataset and Metrics

Two primary benchmark datasets were used for evaluation: PANDORA (3,000 panoramic images, 47 categories) [3] and 360-Indoor (3,334 images, 37 categories) [10]. The PANDORA dataset was split into training (70%, 2,100 images), validation

(15%, 450 images), and test (15%, 450 images) sets. Additionally, our proposed approach was tested on 913 unlabelled images from British Telecom facilities, representing real-world deployment scenarios with object categories outside our training distribution. Performance is measured using mAP@50, i.e., mean average precision at IoU thresholds of 0.5 and 0.5-0.95, where IoU measures the bounding boxes overlap:

$$IoU = \frac{Area(B_{pred} \cap B_{gt})}{Area(B_{pred} \cup B_{gt})} \quad (12)$$

where B_{pred} is the predicted bounding box and B_{gt} is the ground truth. For a single class Average Precision (AP) is computed as the area under the precision-recall curve:

$$AP = \Sigma(R_n - R_{n-1}) \times P_n \quad (13)$$

where P_n and R_n are the precision and recall at the nth threshold. We also employed explainability metrics like percentage of detections with explanations.

B. Comparative analysis

Without using FLS and VLM the object detector identifies objects within its training distribution; however, it's unable to identify everything correctly and also has a lot of duplicate detection issues as shown in Fig. 5 where it detects the same sink twice making the total detections of sinks 4 instead of 3 which is the correct amount, it also detects a cup and bottle that are not there.

Fig. 5. Detections Before Explainable AI enhancement.

By using the FLS to assess the image and VLM to validate those detections it is able to refine the final output to be more correct and increase the certainty of the detected objects as shown in Fig. 6 where it removed the extra detection of the sink, removed the bottle and cup as they are not there and boosted the confidence score of everything that actually exists, while finding 2 extra detections but failing to place their bounding boxes correctly.

Fig. 6. Detections After Explainable AI enhancement.

C. Quantitative Results

Table II presents our performance comparison with state-of-the-art methods. The proposed approach using fuzzy logic rule based system achieves 48.4% mAP on PANDORA, representing a 113% improvement over the previous best result and achieves 51% mAP on 360-Indoor dataset, representing 124.67% improvement over the previous state-of-the-art. The ablation study reveals that fine-tuning alone provides substantial gains ~30% mAP, adding the VLM as a validation engine adds 3-4% mAP while providing crucial explainability with fuzzy logic contributing the final 5.2% mAP.

D. Qualitative Analysis

The system excels at structural elements (walls, floors, ceilings) achieving 65-75% AP, and well-defined objects (TVs, beds, sofas) at 60-70% AP. Challenges remain with small objects (15-40% AP) and thin objects like cables (20-30% AP), primarily due to resolution limitations in extreme distortion regions.

TABLE II. PERFORMANCE COMPARISON ON BENCHMARK DATASETS

Methods	*Pandora mAP@50*	*360-indoor mAP@50*	*Parameters*
Sphere-SSD [7]	12.0%	-	35M
Reprojection R-CNN [8]	16.6%	-	42M
R-CenterNet [3]	22.7%	-	40M
FPN [10]	-	33.6%	-
YOLOv9 (baseline) [13]	8.1%	10.3%	69M
YOLOv9 (fine-tuned) [13]	40.0%	42.1%	57M
Our Method (w/o fuzzy)	43.2%	45.8%	57M
Our Method (complete with fuzzy)	**48.4%**	**51.0%**	57M

Industrial testing on BT's dataset revealed remarkable generalization. The system successfully identified server racks, network equipment, and safety devices never seen during training and managed to reach a mAP of 37.6% on the given classes. The VLM's ability to recognize objects based on visual appearance rather than learned categories proved invaluable. When encountering a fire extinguisher (not in training data), the system correctly identified it as "cylindrical red safety equipment, likely fire suppression device."

E. Types of Explanations

Our system provides three levels of explanations:

- Detection-level: Each bounding box includes specific reasoning. "Chair detected with 72% confidence (boosted from 45% due to clear four-leg structure despite partial table occlusion)."
- Image-level: Overall scene understanding. "Office environment with moderate lighting, high ceiling distortion affecting upper detections, confidence adjustments applied accordingly."
- Error-level: Clear failure explanations. "Possible window missed at θ=45°, φ=30° due to reflection confusion with actual window at θ=50°."

The fuzzy logic system's linguistic output proves crucial for human operators. ***Instead of opaque numerical adjustments, operators see "High distortion zone - relaxed thresholds" or "Excellent quality - strict validation applied." This transparency builds trust and enables informed parameter tuning***.

In one example, the system processed a severely underlit server room. The fuzzy logic identified "very poor lighting, high difficulty" and adjusted thresholds accordingly. The VLM, informed of these conditions, correctly validated equipment detections that would typically be rejected, explaining "server rack confirmed despite low visibility consistent with expected server room layout under emergency lighting.". Analysis of VLM explanations reveals sophisticated reasoning:

- Geometric understanding: "Person detection removed, impossible vertical orientation suggests misclassified pillar"
- Contextual validation: "Keyboard confirmed near monitor despite low confidence, typical workspace arrangement"
- Distortion awareness: "Ceiling light validated despite extreme elongation, expected distortion at φ=85°"

V. Conclusions and Future Work

This paper presented a framework for explainable object detection in 360° panoramic images through integrated fuzzy logic and vision-language models. We addressed the fundamental challenge that existing methods suffer from both poor performance (22-34% mAP on benchmark datasets) and complete lack of explainability, making them unsuitable for industrial deployment. Our approach combines three complementary technologies: a YOLOv9e detector fine-tuned specifically for equirectangular projections with distortion-aware parameter optimization, a fuzzy logic system that evaluates five image quality dimensions and generates linguistic assessments guiding adaptive processing, and a vision-language model that validates detections through semantic reasoning while providing natural language explanations. The framework achieves 48.4% mAP@50 on PANDORA (113% improvement over state-of-the-art R-CenterNet) and 51.04% mAP@50 on 360-Indoor (124.67% improvement over baseline), establishing new performance benchmarks while providing complete explainability through linguistic summarization.

In our future work, we intend to explore Type-2 fuzzy logic systems to model higher-order uncertainties [14], [15]. We also plan to address the VLM bounding box challenge through integration of specialized grounding models like Grounding DINO that explicitly connect language descriptions to precise spatial locations, enabling complete end-to-end detection purely from semantic understanding [16]. The success of fuzzy logic in handling panoramic uncertainties demonstrates that symbolic AI approaches remain essential for transparent decision-making representing a promising direction for developing trustworthy computer vision systems in safety-critical applications.

References

[1] B. Coors, A. P. Condurache, and A. Geiger, "SphereNet: Learning spherical representations for detection and classification in omnidirectional images," in Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018, pp. 518-533.

[2] Q. Zhao, C. Zhu, F. Dai, Y. Ma, G. Jin, and Y. Zhang, "Distortion-aware CNNs for spherical images," in Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018, pp. 1198-1204.

[3] H. Xu, Y. Zhang, B. Zhou, L. Wang, X. Yao, G. Meng, G. Pang, L. Sun, M. Li, C. Pan, and Y. Hu, "PANDORA: A panoramic detection dataset for object with orientation," in Proceedings of the European Conference on Computer Vision, Switzerland, 2022, pp. 695-712.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.

[5] K. -H. Wang and S. -H. Lai, "Object Detection in Curved Space for 360-Degree Camera," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 3642-3646.

[6] R. Cong, K. Huang, J. Lei, Y. Zhao, Q. Huang and S. Kwong, "Multi-Projection Fusion and Refinement Network for Salient Object Detection in 360° Omnidirectional Image," in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 7, pp. 9495-9507, 2023.

[7] M. Cao, S. Ikehata and K. Aizawa, "Field-of-View IoU for Object Detection in 360° Images," in IEEE Transactions on Image Processing, vol. 34, pp. 6813-6822, 2023.

[8] P. Zhao, A. You, Y. Zhang, J. Liu, K. Bian, & Y. Tong, " Spherical Criteria for Fast and Accurate 360° Object Detection," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12959-12966.

[9] T. S. Cohen, M. Geiger, J. Köhler, and M. Welling, "Spherical CNNs," in Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 2018.

[10] S. H. Chou, C. Sun, W. Y. Chang, W. T. Hsu, M. Sun, and J. Fu, "360-Indoor: Towards learning real-world objects in 360° indoor equirectangular images," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 2020, pp. 845-853.

[11] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018, pp. 707-722.

[12] 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, Las Vegas, NV, USA, 2016, pp. 779-788.

[13] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," in Proceedings of the European Conference on Computer Vision, Switzerland, 2024.

[14] H. Hagras, "A hierarchical type-2 fuzzy logic control architecture for autonomous mobile robots," IEEE Transactions on Fuzzy Systems, vol. 12, no. 4, pp. 524-539, 2004.

[15] C. Wagner and H. Hagras, "Toward general type-2 fuzzy logic systems based on zSlices," IEEE Transactions on Fuzzy Systems, vol. 18, no. 4, pp. 637-660, 2010.

[16] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in Proceedings of the European Conference on Computer Vision, Switzerland, 2024, pp. 38-55.

[17] V. Callaghan, M. Colley, H. Hagras, J. Chin, F. Doctor, G. Clarke, "Programming iSpaces: A tale of two paradigms", "*Intelligent Spaces: The Application of Pervasive ICT", (Eds: A. Steventon, S. Wright),* Springer-Verlag, Chapter 24, pp.389- 421, December 2005.

[18] K. Almohammadi, H. Hagras, B. Yao, A. Alzahrani, D.Alghazzawi, G. Aldabbagh,"A Type-2 Fuzzy Logic Recommendation System for Adaptive Teaching" Soft Computing, Vol.21, No.4. pp.965-967, 2017.

[19] F. Doctor, H. Hagras, "Neuro type-2 fuzzy based method for decision making". US Patent 8515884.

[20] N. Sahab, H. Hagras, "Adaptive Non-singleton Type-2 Fuzzy Logic Systems: A Way Forward for Handling Numerical Uncertainties in Real World Applications", IJCCC, Vol. 5, No. 3, pp. 503-529, 2011.