AirEyeSeg: Teacher-Student Insights into Robust Fisheye UAV Detection

Zhenyue Gu[a], Benedikt Kolbeinsson[b] and Krystian Mikolajczyk[c]
Department of Electrical and Electronic Engineering, Imperial College London, London, U.K.

Keywords: Object Detection, Unmanned Aerial Vehicles (UAVs), Fisheye Lenses.

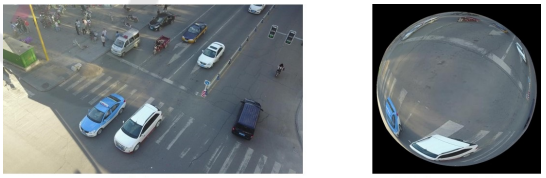
Abstract: Accurate obstacle detection in Unmanned Aerial Vehicles (UAVs) using fisheye lenses is challenged by image distortions. While advanced algorithms like Fast Region-Based Convolutional Neural Network (Fast R-CNN), Spatial Pyramid Pooling-Net (SPP-Net), and You Only Look Once (YOLO) are proficient with standard images, they underperform on fisheye images due to serious distortions. We introduce a real-time fisheye object detection system for UAVs, underpinned by specialized fisheye datasets. Our contributions encompass the creation of UAV-centric fisheye datasets, a distillation-based (also termed Teacher-Student) training method, and AirEyeSeg, a pioneering fisheye detector. AirEyeSeg achieved a Mask(mAP50) of 88.6% for cars on the combined Visdorone and UAVid datasets and 84.5% for people on the SEE dataset using the Box(P) metric. Our results demonstrate AirEyeSeg's superiority over traditional detectors and validate our Teacher-Student training approach, setting a benchmark in fisheye-lensed UAV object detection. The code is available at https://github.com/Zane-Gu/AirEyeSeg.

1 INTRODUCTION

UAVs, commonly known as drones, have transitioned from their initial military applications to diverse civilian sectors, including agriculture, surveillance, topographical mapping, and logistics (Liu et al., 2023). This evolution is attributed to technological advancements and decreasing costs, making UAVs indispensable for both industrial and recreational purposes.

A technique pivotal to the functionality of UAV systems is object detection (Jia et al., 2023). This technique is essential for identifying the category and position of objects of interest within images, thereby providing comprehensive environmental information crucial for scene analysis in computer vision. Given the dynamic and unpredictable nature of aerial environments, it is imperative for UAVs to possess swift and accurate object detection capabilities to ensure safety and avert potential collisions. Dynamic objects such as people, cars etc., are of particular interest and require real time updates due to their changing locations in addition to the moving UAV.

Despite the plethora of proficient object detection algorithms, such as Fast R-CNN (Girshick, 2015), SPP-Net (Purkait et al., 2017), and YOLO (Redmon et al., 2016), which are adept at identifying obstacles in standard images, challenges persist when dealing with images captured through fisheye lenses. The expansive field of view offered by fisheye lenses induces pronounced image distortions, especially near the edges, thereby complicating the task of object delineation (see Figure 1).



(a) Original Image (b) Distorted Image

Figure 1: Example of pronounced distortions in a fisheye image, particularly near the edges.

Addressing the challenges inherent in fisheye image processing proves to be a complicated endeavor, primarily due to the lack of comprehensive datasets specific to fisheye lenses. Established datasets such as Common Objects in Context (COCO) (Lin et al., 2014), ImageNet (Deng et al., 2009), and Visual Object Classes (VOC) (Everingham et al., 2010) do not encompass UAV-centric views, rendering them suboptimal for the development of fisheye object detectors. Moreover, the limited variations in distortion present in current datasets highlight the neces-

[a] https://orcid.org/0000-0002-5182-1226
[b] https://orcid.org/0009-0004-0333-6308
[c] https://orcid.org/0000-0003-0726-9187

Gu, Z., Kolbeinsson, B. and Mikolajczyk, K.
AirEyeSeg: Teacher-Student Insights into Robust Fisheye UAV Detection.
DOI: 10.5220/0012388600003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 546-557
ISBN: 978-989-758-684-2; ISSN: 2184-4313

sity for enriched datasets featuring diverse types of distortions. Prior to initiating the training of our fisheye-adapted detectors, it is imperative to recognize the limitations of traditional training methodologies, which are heavily reliant on original ground truth labels. These methodologies frequently fail to enhance detector generalization, particularly in the common absence of ground truth segmentation. To surmount these challenges, we propose the integration of knowledge distillation, laying the groundwork for a more adaptive and robust fisheye detector.

To sum up, the pivotal contributions of this research are listed as follows:

- Implemented fisheye datasets to bridge the gap in training detectors for UAVs in the field of computer vision.
- Developed a distillation-based strategy for instance segmentation using a foundation model.
- Introduced a novel application of segmentation-based training for the YOLO framework in aerial contexts.
- Engineered and validated AirEyeSeg, a fisheye detector with robust generalization. Notably, it surpasses the performance of detectors trained via traditional methods and sets a new benchmark over the baseline model, YOLOv8.
- Enhanced wire-specific detector to achieve superior detection rates for a broader array of typical UAV obstacles.

The article is structured as follows: Section 2 delves into the evolution of object detection techniques in computer vision. Section 3 elaborates on our chosen datasets and their curation criteria. Section 4 details our training data procedures, introduces the teacher and student models, and describes dataset-specific training approaches. Section 5 presents and analyzes our experimental results. Finally, Section 6 summarizes our findings and suggests avenues for future research.

2 RELATED WORK

Object detection in computer vision has undergone significant evolution, marked by three pivotal milestones: Traditional Detectors, CNN-Based Two-Stage Detectors, and CNN-Based One-Stage Detectors (Zou et al., 2023).

The 1990s heralded the dawn of early computer vision techniques. During this period, algorithms heavily depended on handcrafted features, primarily due to the lack of powerful image representation methods (Viola and Jones, 2004). A quintessential model from this era is Deformable Part-Based Model (DPM), which dominated multiple VOC detection challenges and represented the pinnacle of traditional object detection methods. Felzenszwalb et al. introduced DPM in 2008 as an enhancement of the Histogram of Oriented Gradients (HOG) detector (Zhou et al., 2020), emphasizing a ”divide and conquer” strategy in object detection (Felzenszwalb et al., 2008). This methodology was later refined by Girshick, who incorporated ”mixture models” to cater to a broader range of object variations (Kong et al., 2020). While modern object detectors have outperformed DPM in accuracy, they continue to be influenced by its foundational concepts, including mixture models and bounding box regression.

The renaissance of Convolutional Neural Network (CNN) in 2012 (Krizhevsky et al., 2012) ushered in a new era of advanced object detection techniques. R-CNN was unveiled in 2014, merging selective search with CNNs for object detection, albeit with high computational costs (Girshick et al., 2014). To address this, SPP-Net streamlined the computation of convolutional features (Purkait et al., 2017). Fast R-CNN amalgamated the advantages of R-CNN and SPP-Net, boosting both detection speed and precision (Girshick, 2015). However, its dependency on proposal detection spurred the inception of Faster R-CNN, which integrated a Region Proposal Network (RPN) for streamlined region proposals (Girshick, 2015). Successive models, such as Region-based Fully Convolutional Networks (RFCN) (Dai et al., 2016) and Light head R-CNN (Li et al., 2017), further refined the detection process. Feature Pyramid Network (FPN) employed a top-down architecture, improving object localization across various scales and establishing new standards in object detection (Lin et al., 2017a).

Turning to CNN-based one-stage detectors, the YOLO model was groundbreaking, applying a singular neural network to the entire image for concurrent bounding box and probability estimations (Redmon et al., 2016). Single Shot Detector (SSD) augmented detection accuracy using multireference and multiresolution strategies (Liu et al., 2016). To address class imbalance during training, RetinaNet introduced the ”focal loss” function, narrowing the accuracy disparity between one-stage and two-stage detectors (Lin et al., 2017b). CornerNet reimagined detection as keypoint prediction (Law and Deng, 2018), a concept further streamlined by CenterNet (Zhou et al., 2019). Embracing the Transformer architecture, Detection Transformer (DETR) provided end-to-end detection without the need for anchors, with subsequent

models like Deformable DETR further enhancing this methodology (Carion et al., 2020).

In summary, the trajectory of object detection has witnessed transformative shifts, evolving from a reliance on handcrafted features to leveraging advanced neural networks in a quest for enhanced accuracy and efficiency. However, this pursuit becomes particularly intricate when addressing object detection for UAVs equipped with fisheye lenses, which introduce unique distortions. Traditional training methods, while foundational to the field, often prove suboptimal for the specific challenges of fisheye lens distortions. Recognizing this gap, our primary objective emerges: to engineer real-time detectors adeptly optimized for such lenses. To achieve this, we anchor our solution in the knowledge distillation paradigm (Gou et al., 2021), promising enhanced generalization and performance, and harness the rapid inference capabilities of the renowned YOLO framework, a CNN-based one-stage detector.

3 EXPERIMENTAL DATASETS

Considering the challenges previously emphasized with available datasets, we integrated existing UAV datasets to optimize the training of our object detectors. This section offers an overview of our selected datasets, namely Visdrone (Zhu et al., 2021), UAVid (Lyu et al., 2020), Drone Depth and Obstacle Segmentation (DDOS) (Kolbeinsson and Mikolajczyk, 2023), and the genuine fisheye dataset, SEE[1] data. Figure 2 showcases samples of these datasets. Subsequent to this overview, we elucidate the specifics related to these datasets and further explore the data processing techniques adopted.

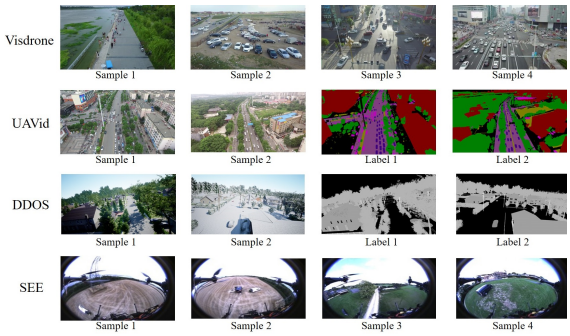


Figure 2: Selected samples from experimental datasets.

The VisDrone2019 dataset (Zhu et al., 2021), encompasses 288 video clips (261,908 frames) and 10,209 static images. These were captured using a range of drone-mounted cameras across 14 distinct cities in China, spanning both urban and rural areas. The dataset captures a diverse array of objects, including pedestrians, vehicles, and bicycles in both sparse and densely populated regions. Our focus is on the VisDrone-VID subset, containing 79 sequences (33,366 frames) at 1344 x 756 resolution. To fit our training emphasis on a resolution of 1280, we doubled the original resolution prior to fisheye transformation. This step ensures that, post-transformation, the fisheye images maintain a high level of detail, vital for our model's training. However, it's worth noting that VisDrone2019 provides ground truth only for object detection, not instance segmentation. To circumvent this limitation, we employed the Segment Anything Model (SAM) (Kirillov et al., 2023) (Subsection 4.2) to produce segmentation masks as ground truth, subsequently facilitating our training process.

The UAVid dataset offers 4K high-resolution UAV imagery centered on urban scenes, consisting of 42 video sequences with 420 labeled images spanning eight classes, including Background clutter, Building, Road, Tree, Low vegetation, Moving car, Static car, and Human (Lyu et al., 2020). Designed for per-pixel semantic labeling, its evaluation relies on the mean Intersection over Union (mIoU) metric. Due to its native 4K resolution, there's no need to upscale before fisheye conversion, preserving the integrity and details of training samples.

The DDOS dataset, also referred to as the wire dataset, contains aerial drone images generated via Airsim[2] simulator, each annotated with pixel-wise segmentation masks. Designed to advance depth estimation and obstacle segmentation research, it encompasses roughly 200GB, featuring frames from Neighbourhood (urban and residential) and Park (natural landscapes) environments. With a resolution of 1280 x 720, it consists of ten classes, including ultra-thin (wires), thin structures, small mesh, large mesh, trees, buildings, vehicles, people, animals, and others. The dataset chiefly supports wire detector training, predominantly from its 25,300-image Neighbourhood segment. For evaluations, models are trained using original DDOS labels, negating the need for resolution tweaks.

The SEE dataset, captured using a genuine fisheye UAV, serves exclusively to evaluate our model's performance. It consists of 30,136 frames without any ground truth labels. Thus, similar to the aforementioned datasets, we employed SAM to generate segmentation masks for cars, people, trees, and buildings as ground truth. All frames maintain a resolution of

[1]The SEE dataset will be made available after the final revision of this paper due to double blind review process.

[2]https://microsoft.github.io/AirSim/

1280 x 960.

We have collated four datasets offering a spectrum of views from urban to rural landscapes, essential for advancing fisheye detector research. Renowned datasets such as Pascal VOC, COCO, ImageNet, and FishEye8K (Gochoo et al., 2023) proffer a vast collection of images but fall short in providing UAV-centric perspectives and wire representation—central elements in our investigation. In contrast, our datasets prioritize these critical features. Given their comprehensive nature, a requisite step involves their conversion to a fisheye format, yielding datasets characterized by a spectrum of distortions and redressing prevailing dataset inadequacies. Our training framework focuses on cardinal obstacle classes, including cars (spanning diverse vehicle types), people (aggregating both humans and pedestrians), trees, and buildings. It's worth noting that Visdrone, UAVid, and DDOS support both training and testing phases, however, DDOS distinctively features wire classification while omitting the people category, and the SEE dataset is dedicated exclusively to testing. A comprehensive discussion on the applied processing techniques is scheduled for the subsequent section.

4 AirEyeSeg

In this section, we present our Fisheye object detector for aerial images. Figure 3 illustrates the Teacher-Student training scheme employed to enhance our obstacle detectors, ensuring robust generalization across diverse focal lengths of fisheye lense. Before feeding the chosen standard images into the teacher—a foundation image segmentation model—they are converted to a fisheye format which is discussed in Section 4.1. We then present the segmentation teacher and student models in Section 4.2 and Section 4.3, respectively. We have employed Segment Anything Model (SAM) (Kirillov et al., 2023) as the teacher and YOLOv8 (Jocher et al., 2023) as the student. Section 4.4 presents our training strategies inspired by knowledge distillation framework (Gou et al., 2021). We train various detectors to evaluate and compare their performance in Section 5.

4.1 Fisheye Data Generation

As depicted in Figure 3, the initial phase of developing fisheye detectors necessitates the transformation of standard image datasets into a fisheye format. Subsequently, data augmentation techniques are employed to enhance the dataset, culminating in a comprehensive repository of fisheye images.

4.1.1 Fisheye Distortions

We discuss the methodology for converting standard images, also referred to as conventional images, into fisheye format. The transformation process is based on a method proposed by (Ye et al., 2020), wherein a mapping is established from the fisheye image plane to the conventional image plane. This mapping can be reversed to projects conventional images into the fisheye domain. Conventional images are typically captured by a pinhole camera. The perspective projection associated with this model is articulated in Equation 1. On the other hand, fisheye cameras often employ the equidistant projection model, as described in Equation 2 (Kannala and Brandt, 2006):

$$r_{pinhole} = f \tan\theta, \tag{1}$$

$$r_{equidistance} = f\theta, \tag{2}$$

Here, θ represents the angle between the principal axis and the incoming ray, r denotes the distance from the image point to the principal point, and f is the focal length.

Both conventional and fisheye images can be understood as projections of a hemisphere onto a plane, albeit using different projection models and view angles. The projection model for fisheye images is illustrated in Figure 4.

The intricacies of the geometric imaging model are elaborated upon in (Kannala and Brandt, 2006). Assuming congruent focal lengths for both the perspective and equidistance projections, and setting the maximum viewing angle θ_{max} to 180°, the transformation from the fisheye image point $P_f = (x_f, y_f)$ to the conventional image point $P_c = (x_c, y_c)$ is articulated in Equation 3:

$$r_c = f \tan(r_f / f), \tag{3}$$

In this context, $r_c = \text{sqrt}((x_c - u_{cx})^2 + (y_c - u_{cy})^2)$ represents the distance between the image point P_c and its principal point $U_c = (u_{cx}, u_{cy})$ in the conventional image. Similarly, $r_f = \text{sqrt}((x_f - u_{fx})^2 + (y_f - u_{fy})^2)$ denotes the corresponding distance in the fisheye image between the image point P_f and its principal point $U_f = (u_{fx}, u_{fy})$.

The transformation relationship encapsulated in Equation 3 is intrinsically modulated by the focal length f. By instituting a foundational focal length f_0, the fisheye camera model is approximated to span a hemispherical visual domain. Each image, paired with its respective annotation from existing standard image datasets, is subjected to this transformation, employing the aforementioned mapping function, thereby facilitating the synthesis of fisheye

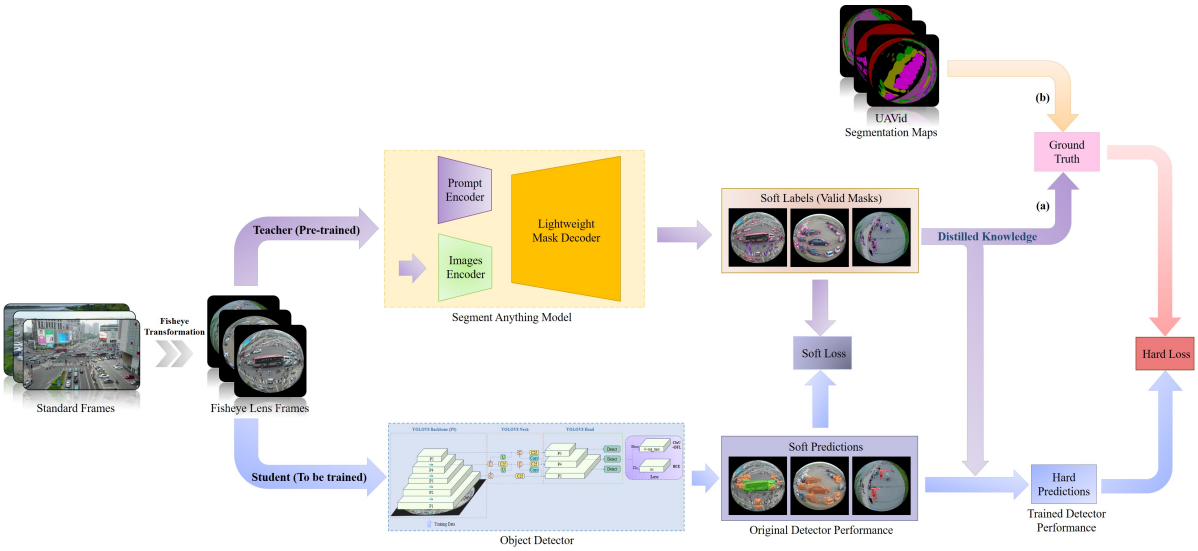


Figure 3: Teacher-Student training scheme for dual fisheye object detectors: A computer vision foundation model is the teacher, guiding object detectors as students to enhance detection and segmentation for obstacle avoidance: (a) AirEyeSeg training; (b) AirEyeGT training. The soft loss quantifies discrepancies between soft label and soft prediction probability distributions, while the hard loss evaluates differences between ground truth and the student model's hard predictions.

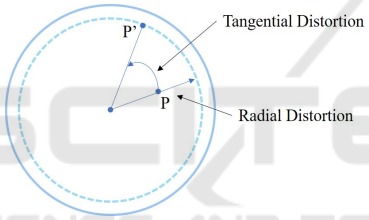


Figure 4: Illustration of fisheye distortion model. P is the perspective imaging point and $P^{'}$ is the fisheye imaging point. Radial distortion indicates deviation along the imaging radius, while tangential distortion denotes displacement along the tangential direction of the imaging point (Li et al., 2020).

image datasets. Fisheye images are generated from existing standard images, accompanied by their respective segmentation maps, for both both $f = 150$ and $f = 300$, as illustrated in Figure 5. This augmentation strategy is designed to enrich the diversity of fisheye image datasets and to strengthen the generalization capability of detectors across different levels of distortions.

4.1.2 Data Augmentation

To increase the diversity of our fisheye image datasets, data augmentation is crucial, addressing the inherent limitations in training image variance. We incorporate horizontal flipping, rotation, and random cropping to the standard images. Horizontal flipping mirrors images along their vertical axis, enhancing orientation diversity. Rotational adjustments vary,

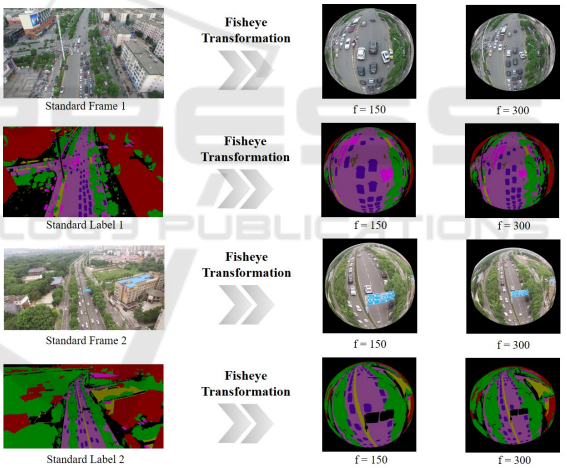


Figure 5: Transformation of original standard frames and segmentation maps into their respective fisheye counterparts.

from pronounced shifts of $\pm 90^{\circ}$ to subtle changes ranging between -15° and $+15^{\circ}$. Concurrently, random cropping, truncating up to 20% of the image, refines the model's capability to detect partially obscured objects.

4.1.3 Post-Processing

Upon securing segmentation maps from SAM (Kirillov et al., 2023), it is necessary to transform these masks—which include both SAM-generated masks and original dataset ground truth labels—into polygo-

nal representations, a prerequisite for YOLO training given its inability to directly interpret segmentation maps. The process initiates by distinguishing each object or class in an image via unique segmentation map colors. For each identified class, a binary mask is created, isolating specific class instances. Differentiating multiple instances of the same class in an image involves connected components analysis and assigning unique labels to interconnected regions. Each instance is then transmuted into polygons based on binary mask contours. Importantly, to mitigate potential representation inaccuracies, polygons are initially generated for all object classes output by SAM, with subsequent filtering to retain only those relevant to our study. This method effectively curtails conversion discrepancies. The resultant polygons are cataloged in the COCO format, as illustrated in Figure 6, encompassing not only polygon coordinates but also integrating vital metadata such as bounding boxes and area.

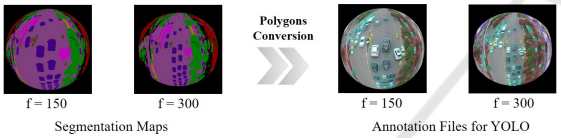


Figure 6: Examples of conversion from segmentation masks to annotation polygons.

4.2 Teacher: Image Segmentation Foundation Model

Our methodology integrates SAM, a segmentation model (Kirillov et al., 2023), as a foundational instructive entity throughout the comprehensive training process. Distinguished for its proficiency in segmenting a diverse array of objects, SAM serves as the guiding force for the object detectors, which function as students within the training framework. For completeness of this paper we briefly overview the architecture of SAM with its associated theoretical underpinnings but the details can be found in (Kirillov et al., 2023).

Figure 7 presents SAM, which comprises three primary components: an image encoder, a versatile prompt encoder, and an efficient mask decoder. Its image encoder system integrates a Vision Transformer (ViT) (Dosovitskiy et al., 2020), which is pre-trained using Masked Autoencoders (MAEs) (He et al., 2022) and tailored for high-resolution inputs (Li et al., 2022). This encoder processes each image once, offering the possibility of pre-prompt application. The framework of prompt encoder distinguishes between two types of prompts: sparse (such as points, boxes, and text) and dense (like masks). Positional encodings, in conjunction with specific learned embeddings, represent points and boxes. A renowned text encoder from Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) manages free-form text. For dense prompts, particularly masks, convolutional embeddings are employed, which are element-wise merged with the image embedding. The mask decoder converts the amalgamated image and prompt embeddings, together with an output token, into a mask. The design, inspired by (Carion et al., 2020), incorporates a refined transformer decoder block followed by a dynamic mask prediction head. This decoder block facilitates bi-directional cross-attention between prompts and image embeddings. After two block operations, the image embedding undergoes upsampling. Subsequently, an Multilayer Perceptron (MLP) transforms the output token into a dynamic linear classifier, which ascertains the mask foreground probability for each image pixel.

4.3 Student: Efficient Object Detector

We have chosen YOLOv8 (Jocher et al., 2023), the most advanced iteration of the YOLO series, to function as our object detector, guided by the teacher model. As depicted in Figure 8, YOLOv8 architecture encompasses four principal components: the backbone, neck, head, and loss. While YOLOv8 is grounded in the foundational principles of YOLOv5 (Jocher et al., 2022), it incorporates several notable enhancements.

YOLOv8 adapts the backbone of YOLOv5, introducing modifications to the Cross-Stage-Partial-connection (CSP) Layer, now termed the C2f module, which divides the feature map into two segments. The first segment undergoes convolution, while the second is concatenated with the first's output, enhancing the learning capacity of CNNs and amalgamating high-level features with contextual data to boost detection accuracy while reducing computational cost.

The neck of YOLOv8 employs multi-scale feature fusion, utilizing the Feature Pyramid Network (FPN) (Lin et al., 2017a) and Path Aggregation Network (PAN) (Liu et al., 2018) architectures to integrate features from various network layers (Ju and Cai, 2023). The FPN upsamples to enhance the bottom feature map, and the PAN downsamples to strengthen the top feature map, ensuring precise predictions across diverse image sizes and optimizing computational efficiency by incorporating the FP-PAN and omitting convolution operations during upsampling.

Contrasting to YOLOv5's coupled head, YOLOv8 has an anchor-free, decoupled head (Terven and Cordova-Esparza, 2023), allowing independent pro-

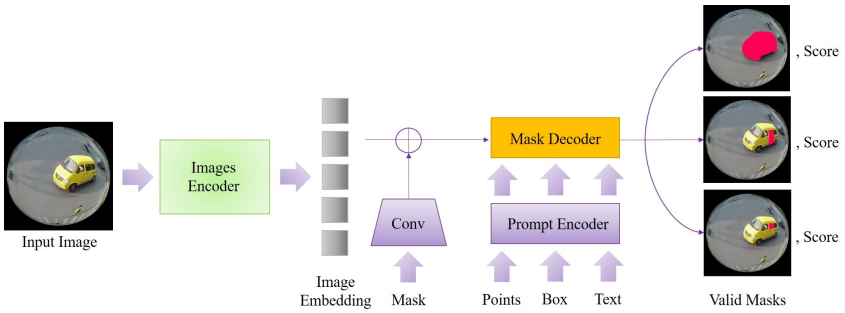


Figure 7: The architecture of SAM. A heavyweight image encoder produces an image embedding, which can be swiftly queried using diverse input prompts to generate object masks at near real-time speeds. For prompts that correspond to multiple objects, SAM is capable of producing several valid masks along with their associated confidence scores.

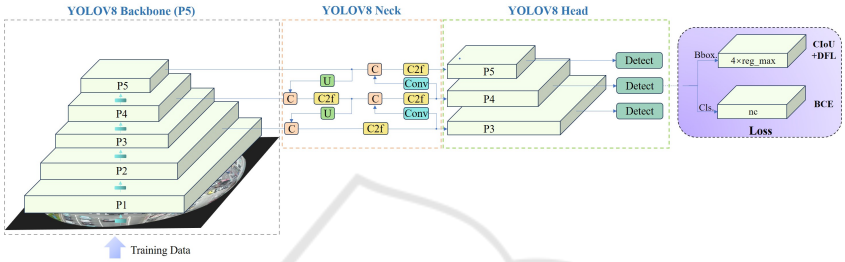


Figure 8: The architecture of YOLOv8. The architecture primarily comprises four components: Backbone, Neck, Head, and Loss.

cessing of objectness, classification, and regression tasks and enhancing overall accuracy by enabling each branch to specialize. The output layer employs the sigmoid function for the objectness score and the softmax function for class probabilities, indicating the likelihood of a bounding box containing an object and the probability of objects correlating with each potential class, respectively. The model incorporates the Complete IoU (CIoU) and Distribute Focal Loss (DFL) as loss functions to address bounding box loss. Together, these loss functions optimize the model's ability to detect objects, particularly enhancing accuracy with smaller objects.

In our approach, we employed YOLOv8's segmentation variant, YOLOv8-Seg (Jocher et al., 2023). This model is based on the CSPDarknet53 extractor and features a unique C2f module, diverging from traditional YOLO neck designs. Following this are two segmentation heads and detection modules mirroring those in YOLOv8. Notably, YOLOv8-Seg sets new benchmarks in object detection and semantic segmentation without sacrificing speed or efficiency.

4.4 Training Strategies

As underscored earlier, the objective of this study is to develop object detectors specifically designed for fisheye lenses, with the aim of creating multiple detectors utilizing diverse datasets for comparative evaluation. In this endeavor, two distinct training strategies were employed on different datasets: two detectors were developed using distillation-based training on the VisDrone and UAVid datasets to identify common UAV obstacles encountered in real-world scenarios, such as cars, people, trees, and buildings. Conversely, subsequent detectors employing a traditional training strategy primarily targeted the detection of wires, cars, buildings, and trees on the DDOS dataset.

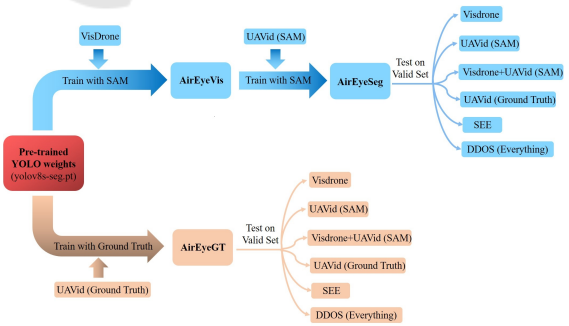


Figure 9: Teacher-Student training on Visdrone and UAVid datasets: AirEyeSeg is trained utilizing SAM-annotated labels as ground truth, whereas AirEyeGT employs real ground truth labels.

4.4.1 Teacher-Student Training

As depicted in Figure 9, the original YOLOv8 segmentation weight, yolov8s-seg.pt, was fine-tuned utilizing fisheye lens images extracted from the Visdrone and UAVid datasets. A salient distinction emerges between these detectors; AirEyeSeg was trained employing SAM-generated masks for both its validation and test sets, while AirEyeGT utilized real ground truth labels from the UAVid dataset for the corresponding sets. It is noteworthy that AirEyeGT was exclusively trained on the UAVid dataset—a decision predicated on the absence of instance segmentation masks in the Visdrone dataset, as discussed in Section 3. This training approach was adopted to enable a comparative analysis between the two detectors, thereby elucidating the efficacy of SAM in real-time segmentation prediction.

4.4.2 Ground Truth Based Training

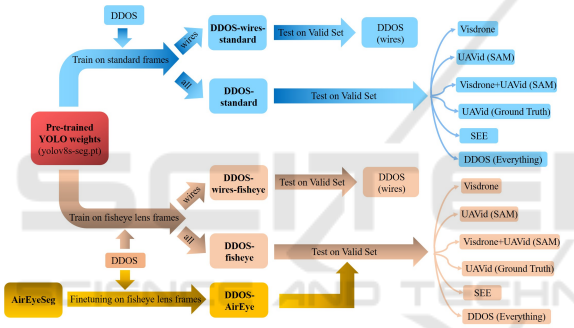


Figure 10: Traditional training on DDOS dataset; Five detectors were trained; two on standard frames (distinguishing between wires-only and all obstacles), and two on fisheye lens frames with similar distinctions. Additionally, the weights of AirEyeSeg were fine-tuned for comparison. All training phases utilized real ground truth labels.

A traditional training strategy was employed for the DDOS dataset as presented in Figure 10. Within this framework, five detectors were meticulously trained using real ground-truth labels across training, validation, and testing sets, enabling a thorough comparison of detection performance. This also provided insight into contrasting this traditional training performance with our devised Teacher-Student approach. In addition, two detectors were trained using standard non-distorted frames from the DDOS dataset. The first, denoted as DDOS-wires-standard, was tailored specifically for wire detection, while the second, termed DDOS-standard, was engineered to identify a variety of obstacles as previously mentioned. In the case of detectors trained on fisheye lens frames, the training approach mirrored that of the standard frame detectors, leading to the development of DDOS-wires-fisheye and DDOS-fisheye detectors, respectively. Additionally, the weight of AirEyeSeg was fine-tuned on fisheye lens frames, establishing a comparative performance benchmark against other detectors.

5 RESULTS ANALYSIS

In this section, we introduce the metrics for our experimental evaluation, followed by an analysis and visualization of our detectors' performance across all datasets.

To evaluate the segmentation performance of our trained YOLOv8-based detectors we use Precision (P) (Euzenat, 2007) and mean Average Precision (mAP) (Everingham et al., 2010). In the subsequent analysis of experimental results, the performance of the trained detectors is evaluated using mAP50, which is determined with an Intersection over Union (IoU) threshold of 0.50.

5.1 Comparative Analysis of Developed Detectors

The primary aim of this subsection is to rigorously evaluate the performance of the models across all datasets, emphasizing the Mask (mAP50) and Box (P) metrics. Initially, we assess the developed detectors on the test sets of our training datasets, including Vidrone, UAVid, and DDOS. Subsequently, to further objectively evaluate their performance, we tested the detectors on the SEE dataset, which is unfamiliar to all detectors and is only used for testing.

5.1.1 Evaluation on VisDrone, UAVid and DDOS

Figure 11 compares the performance of five detectors against YOLOv8 on the combined Visdrone and UAVid datasets, showcasing the effectiveness of our Teacher-Student training approach compared to the other detectors trained in a traditional ground truth supervised approach. Notably, AirEyeGT and AirEyeSeg, trained with distilled knowledge, exhibit superior mAP50 values. Specifically, in the cars category, AirEyeSeg achieved an mAP50 of 88.6%, followed by AirEyeGT at 85.6%. The similar performance of the knowledge-distilled detectors in Mask(mAP50) and Box(P) indicates the high reliability of the SAM-generated annotations. However, AirEyeSeg consistently outperforms AirEyeGT across most classes due to its extended training dataset.

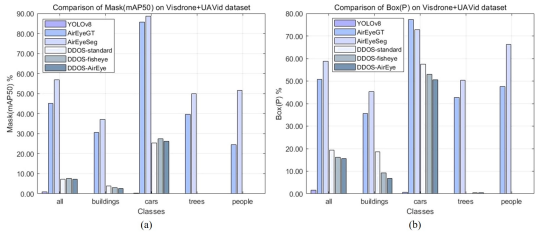


Figure 11: Comparative analysis of metrics across six algorithms on the combined Visdrone and UAVid Datasets: (a) Mask(mAP50); (b) Box(P). The evaluated algorithms comprise the original YOLOv8, AirEyeGT, AirEyeSeg, DDOS-standard, DDOS-fisheye, and DDOS-AirEye. See Figure 9 and 10 for how different detectors were trained.

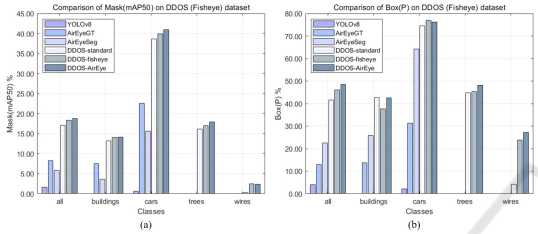


Figure 12: Comparative Analysis of Metrics Across Six Algorithms on the DDOS dataset: (a) Mask(mAP50); (b) Box(P). The evaluated algorithms comprise the original YOLOv8, AirEyeGT, AirEyeSeg, DDOS-standard, DDOS-fisheye, and DDOS-AirEye. See Figure 9 and 10 for how different detectors were trained.

Figure 12 presents the performance metrics of the six models on the DDOS dataset. As anticipated, DDOS-standard, DDOS-fisheye, and DDOS-AirEye outperform AirEyeGT and AirEyeSeg in both metrics, as they were trained and tested on synthetic DDOS data only, while AirEyeGT and AirEyeSeg were trained on real VisDrone and AUVid datasets. Remarkably, DDOS-AirEye eclipses DDOS-standard and DDOS-fisheye in obstacle identification, registering performance enhancements of 1.7% and 0.5% in Mask(mAP50), and 7% and 2.5% in Box(P), respectively. This performance surge can be attributed to the fine-tuning of DDOS-AirEye based on AirEyeSeg weights, underscoring the broad applicability of our Teacher-Student training paradigm across diverse datasets. The consistent outcomes observed between AirEyeSeg and AirEyeGT further validate the authenticity of teacher-generated annotations.

5.1.2 Evaluation on SEE Dataset

To corroborate our earlier findings, we evaluated six algorithms on the SEE dataset that was captured in different environment and was not used for training the detectors therefore can be considered of different distribution than the training sets.

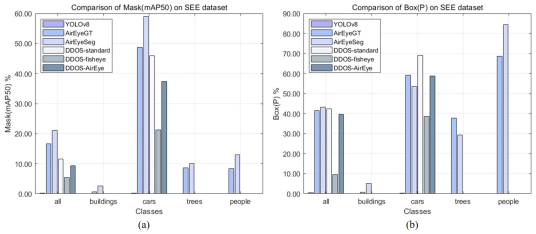


Figure 13: Comparative analysis of metrics across six algorithms on the SEE dataset: (a) Mask(mAP50); (b) Box(P). The evaluated algorithms comprise the original YOLOv8, AirEyeGT, AirEyeSeg, DDOS-standard, DDOS-fisheye, and DDOS-AirEye. See Figure 9 and 10 for how different detectors were trained.

As illustrated in Figure 13, the outcomes align with our previous findings. Notably the top performance is significantly lower than for the other datasets presented in Figure 11 and 12, where the models were trained from different train/test splits of the same datasets. Despite that, using the Mask(mAP50) metric as a reference, detectors that employ the Teacher-Student training paradigm demonstrate superior performance relative to those trained on the DDOS dataset using conventional methods. This superiority is further emphasized by the overall class detection values represented by Box(P). Across both metrics, the performance of the student detectors remains consistent. Notably, AirEyeSeg not only consistently surpasses AirEyeGT in the Mask(mAP50) metric but also outperforms in the overall classes as indicated by the Box(P) metric, showcasing its enhanced accuracy in obstacle detection. Additionally, AirEyeSeg stands out in detecting the people class, achieving an 84.5% accuracy rate as denoted by the Box(P) metric. These findings underscore the potency of the Teacher-Student training approach for fisheye detectors.

In addition, when assessed across both metrics, DDOS-AirEye consistently surpasses the DDOS-fisheye detector, further emphasizing the superior adaptability of our student detector. Importantly, the DDOS-standard detector also demonstrates exceptional performance, rivaling AirEyeSeg in the Box(P) metric. Such performance can be attributed to the minimal distortion present in the SEE dataset. Considering that the SEE dataset more accurately reflects standard frames compared to training datasets with varied distortion levels, the DDOS-standard detector attains noteworthy test results.

5.2 Detection Visualizations

Here, we highlight visual results from our top-performing detectors, AirEyeSeg and DDOS-AirEye, compared with the original YOLOv8, to under-

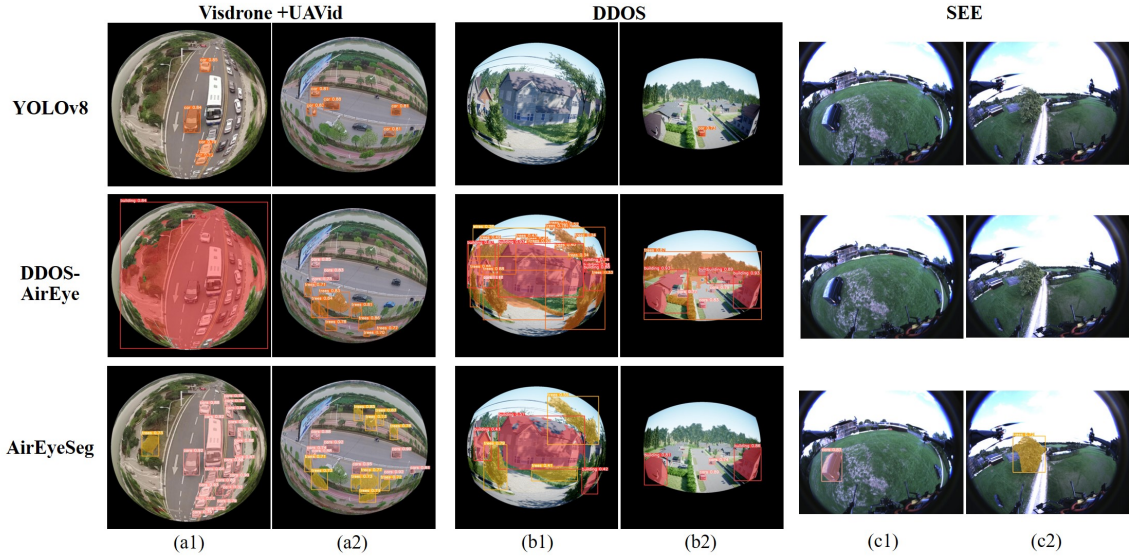


Figure 14: Comparative visualization of obstacle detection outcomes for developed algorithms vs. YOLOv8 across all datasets: (a) samples from Visdrone and UAVid datasets; (b) samples from DDOS dataset; (c) samples from SEE dataset.

score our advancements. These visualizations vividly demonstrate the precision of developed detectors in identifying and delineating common obstacles and wires across various datasets.

We visually showcase our detectors' ability to identify obstacles like cars, trees, buildings, and wires, each annotated with a class label and confidence score. These visuals, as seen in Figure 14, not only highlight our algorithms' accuracy but also echo our earlier quantitative findings.

6 CONCLUSIONS

Our rigorous analysis underscores the potentials of the foundation models in Teacher-Student training approach for object detection pertinent to fisheye UAVs, with AirEyeSeg exemplifying notable superiority over conventional detectors and evidencing the robustness of teacher-generated labels. However, the nuanced and slender characteristics of wires introduce persistent challenges in detection. As we delineate future trajectories, pivotal research avenues include: 1) Probing advanced YOLO variants, building on YOLO's foundational success in our study and aiming to harness the innovations of its latest iterations; 2) Refining the transformation from segmentation masks to polygons, a critical step to enhance the granularity and precision of detector evaluations; 3) Expanding the scope of Teacher-Student training methodologies, leveraging their demonstrated efficacy across diverse realms within computer vision.

ACKNOWLEDGEMENTS

This work was supported by Innovate UK project 10021134.

REFERENCES

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Dai, J., Li, Y., He, K., and Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural inf. proc. systems*, 29.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Euzenat, J. (2007). Semantic precision and recall for ontology alignment evaluation. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 348–353. AAAI Press.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338.

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable

part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Gochoo, M., Otgonbold, M.-E., Ganbold, E., Hsieh, J.-W., Chang, M.-C., Chen, P.-Y., Dorj, B., Al Jassmi, H., Batnasan, G., Alnajjar, F., Abduljabbar, M., and Lin, F.-P. (2023). Fisheye8k: A benchmark and dataset for fisheye camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Jia, X., Tong, Y., Qiao, H., Li, M., Tong, J., and Liang, B. (2023). Fast and accurate object detector for autonomous driving based on improved yolov5. *Scientific reports*, 13(1):1–13.

Jocher, G., Chaurasia, A., Qiu, J., and Ultralytics (2023). Ultralytics yolov8: State-of-the-art model for real-time object detection, segmentation, and classification. https://github.com/ultralytics/ultralytics. Accessed: 2023-08-28.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., Michael, K., Fang, J., Yifu, Z., Wong, C., Montes, D., et al. (2022). ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo*.

Ju, R.-Y. and Cai, W. (2023). Fracture detection in pediatric wrist trauma x-ray images using yolov8 algorithm. *arXiv preprint arXiv:2304.05071*.

Kannala, J. and Brandt, S. S. (2006). A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Kolbeinsson, B. and Mikolajczyk, K. (2023). DDOS: The drone depth and obstacle segmentation dataset. arXiv preprint arXiv:2312.12494.

Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., and Shi, J. (2020). Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Law, H. and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750.

Li, T., Tong, G., Tang, H., Li, B., and Chen, B. (2020). Fisheyedet: A self-study and contour-based object detector in fisheye images. *IEEE Access*, 8:71739–71751.

Li, Y., Mao, H., Girshick, R., and He, K. (2022). Exploring plain vision transformer backbones for object detection. pages 280–296.

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., and Sun, J. (2017). Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017a). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Liu, H., Duan, X., Lou, H., Gu, J., Chen, H., and Bi, L. (2023). Improved gbs-yolov5 algorithm based on yolov5 applied to uav intelligent traffic. *Scientific Reports*, 13(1):9577.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer.

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., and Yang, M. Y. (2020). Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119.

Purkait, P., Zhao, C., and Zach, C. (2017). Spp-net: Deep absolute pose regression with synthetic views. *arXiv preprint arXiv:1712.03452*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. pages 8748–8763.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 779–788.

Terven, J. and Cordova-Esparza, D. (2023). A comprehensive review of yolo: From yolov1 and beyond. arxiv 2023. *arXiv preprint arXiv:2304.00501*.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57:137–154.

Ye, Y., Yang, K., Xiang, K., Wang, J., and Wang, K. (2020). Universal semantic segmentation for fisheye urban driving images. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 648–655. IEEE.

Zhou, W., Gao, S., Zhang, L., and Lou, X. (2020). Histogram of oriented gradients feature extraction from raw bayer pattern images. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 67(5):946–950.

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., and Ling, H. (2021). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399.

Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*.