

Supplementary Materials: Uni-YOLO: Vision-Language Model-Guided YOLO for Robust and Fast Universal Detection in the Open World

Anonymous Authors

1 NOTES ON SUPPLEMENTARY MATERIALS

The supplementary materials provide additional details of the proposed methods and experimental results. The summary of the supplementary materials is as follows:

- Details of the proposed physical model-based Enhancement Module (EHM) are provided in Section 2.
- A joint training method for the proposed EHM with the universal detector Uni-YOLO is provided in Section 3.
- The physical representation of the developed UAV detection platform is provided in Section 4.
- More visual detection results in the real world based on our UAV detection platform are provided in Section 5.

2 THE DETAIL ARCHITECTURE OF THE EHM

Our EHM has a two-branch architecture to enable the estimation of two sets of physical parameters using a single backbone. This efficient architecture enables real-time detection performance.

As shown in Figure 3, the backbone is designed as a dense connection with four convolutional layers to enable the fusion of feature information at multiple scales. We ensure that the output size of each convolutional layer is consistent with the input image by adjusting the size of the convolutional kernel and the size of the zero padding. Then, the multi-scale feature group is divided into two feature branches according to an adaptively weighted architecture, the first feature group mainly contains the dehazing map information, and the second feature group mainly contains the correction map information. The feature pooling and 1D convolution perform the adaptive weighting. Finally, the two weighted features are individually converted into the dehazing map $D_M(x)$ and the correlation map $\gamma_{r,g,b}(x)$ by adjusting the number of channels through convolution layers. Table 7 shows the detailed parameters of our EHM and the corresponding output size.

Table 7: The parameter details of the proposed EHM. We use c to denote the input channel number, k to denote the kernel size, and p to denote the zero padding number.

Layers	Configurations	Output Size
Input	Degraded Images	$h \times w \times 3$
Conv 1	$c = 3, k = 3, p = 1$	$h \times w \times 8$
Conv 2	$c = 8, k = 3, p = 1$	$h \times w \times 8$
Conv 3	$c = 16, k = 3, p = 1$	$h \times w \times 8$
Conv 4	$c = 24, k = 3, p = 1$	$h \times w \times 8$
Conv 5	$c = 32, k = 3, p = 1$	$h \times w \times 3$
Conv 6	$c = 32, k = 3, p = 1$	$h \times w \times 3$

3 THE JOINT TRAINING METHOD FOR EHM

To achieve adaptive detection-friendly enhancement for input images, the proposed EHM uses a joint training method. First, the EHM is sequentially trained using the low-light enhancement dataset and scattering recovery dataset to provide it with initial enhancement capability. It is then trained jointly with the universal detector to achieve adaptive detection-friendly enhancement. The specific steps of the training method are summarised in Algorithm 3.

Algorithm 3 The Training Method for our EHM

Input: Degraded images with their clear reference (img_{de}, img_{re});
Detection images with their annotations ($img, label$).
// Step 1. For pre-training of the EHM(img).
1: **for** k in ($img_{de}(n), img_{re}(n)$) **do**
2: Update the network $EHM(img_{de}(k))$, with the perceptual loss: $Loss_{pc}(EHM(img_{de}(k)), img_{re}(k))$.
3: **end for**
// Step 2. For joint training of the EHM and Uni-YOLO.
4: Constructing integrated model: Uni-YOLO $\Theta(img)$ with EHM.
5: **for** m in ($img(n), label(n)$) **do**
6: Update the network $EHM(img(m))$ and Uni-YOLO $\Theta(img(m))$ using $Loss_{goc}$, $Loss_r$ and $Loss_{con}$.
7: **end for**
Output: The trained EHM for detection-friendly enhancement.

4 THE DEVELOPED UAV PLATFORM

Based on the proposed Uni-YOLO, we developed a UAV system for detecting objects in the open world. Figure 7 shows the physical representation of the UAV detection system. With low energy consumption and long endurance, our UAV detection system can be used in various real-world environmental monitoring tasks.

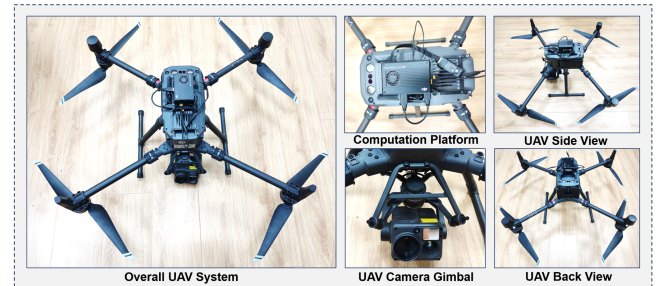


Figure 7: The physical representation of the UAV detection system. The MANIFOLD 2 (with Jetson TX2) is our computation platform for the proposed OARE-based detector.

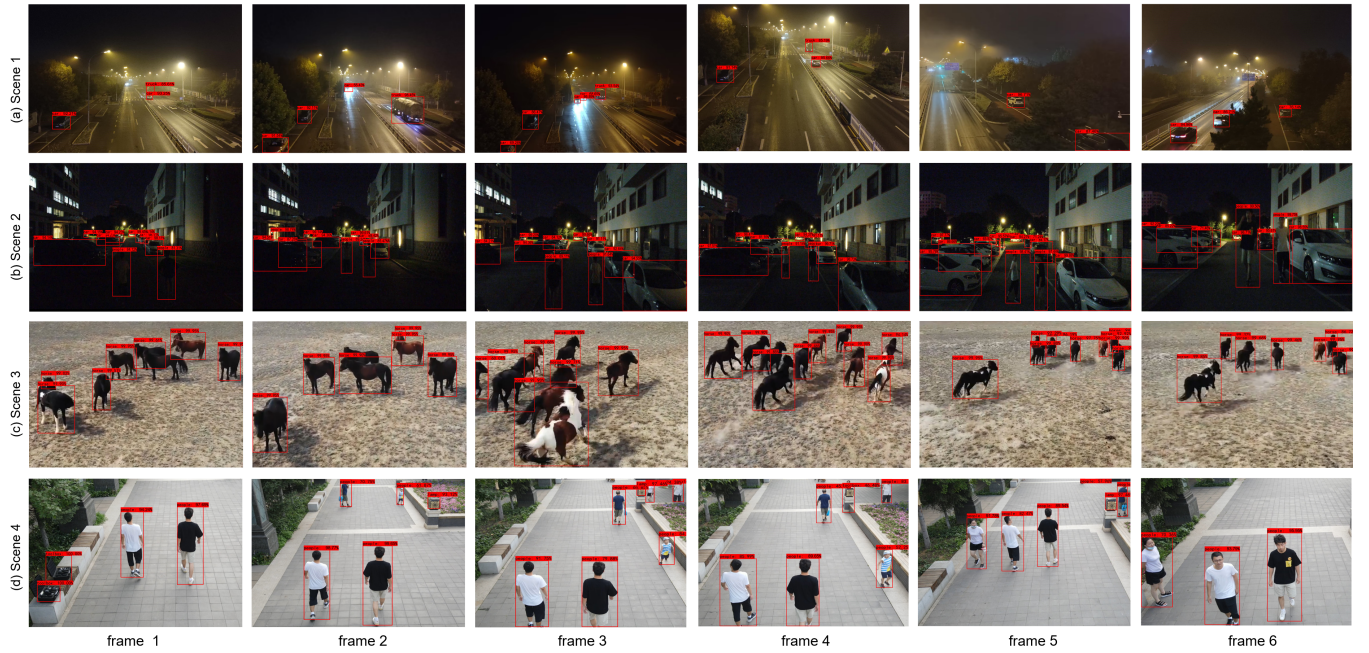


Figure 8: Illustration for the zero-shot detection performance of Uni-YOLO based on the developed UAV in the open world. (a) Low illuminance and hazy conditions. (b) Low illuminance conditions. (c) Natural desert scenes. (d) Nature park scenes. These images show the real-time detection when the UAV is in flight, showing detection results every 5 seconds.

5 THE DETECTION RESULTS BASED ON UAV

We use the developed UAV detection platform to detect objects in the open world. Figure 8 shows some detection results of our UAV-based Uni-YOLO while the UAV is in flight. (a) In low illuminance and hazy conditions, the candidate objects of interest to the user are "car" and "truck". (b) In low illuminance conditions, the candidate objects of interest to the user are "people" and "car". (d) In natural

desert scenes, the candidate object of interest to the user is "horse". (b) In natural park scenes, the candidate objects of interest to the user are "people", "toolbox" and "lamp". The detection results show that our Uni-YOLO can provide robust detection performance in various harsh and normal conditions. In addition, we also provide a video to show the detection performance of Uni-YOLO in a real-world nighttime scene (please refer to "test.mp4").