Attention-Driven Mixing Perception based on Object Detection for UAV Distance Estimation

Duan Yuan[1] , *Yicheng Luo*[1] , *Rihong Yan*[1] *Jiafan Zhuang*[1*,] , *Zhun Fan*[2*]

(1. Department of Electronic Engineering, Shantou University, Shantou, Guangdong;

2. Shenzhen Institute for Advanced Study, UESTC)

Abstract: UAV target distance estimation plays a crucial role in enabling the emergence of collective intelligence for swarm UAVs. The basis for estimating the distance of a UAV target lies in object detection. Although object detection models learn essential information for distance estimation, most research tends to overlook the relationship between object detection and distance estimation. In light of this observation, we propose a novel approach that utilizes self-view attention and cross-view attention mechanisms to enhance object detection perception for accurate distance estimation. To learn the relationship between object detection and distance estimation, we stack the attention mechanism and distance estimation model, training them in an end-to-end manner. We conduct experiments on the UAVDE dataset and demonstrate that our method achieves a significant improvement over the state-of-the-art method PCM (reducing the relative difference from 12.0% to 10.1%. This highlights the effectiveness and superiority of our proposed approach.

Keywords: drone perception, object detection, distance estimation, deep learning.

0 Introduction

Due to shared perception and intelligence navigation, swarm UAVs have a unique advantage in many applications, such as rescue [1], 3d reconstruction [2], and environment exploration [3]. With the development of swarm UAVs, collective intelligence has emerged as a powerful concept, showcasing remarkable flexibility and exploration capabilities. Accurately detecting targets and estimating distances are vital components of collective intelligence, as they provide crucial information to support perception and navigation processes Accurate UAV target detection and distance estimation remain challenging in practice due to several factors. One of the main obstacles is the scarcity of datasets that contain both stereo images and corresponding distances. Most stereo-based distance estimation methods

Received date: 2024-09-30

*** E-mail:** jfzhuang@stu.edu.cn, fanzhun@uestc.edu.cn

rely on dense disparity maps as labels, which are typically annotated using Lidar. However, Lidar is not suitable for UAV targets, making it difficult to obtain reliable distance annotations [4]. Furthermore, UAVs have limited resources, such as power, computational compatibility, and physical size, which constrain the implementation of complex algorithms. To address the lack of suitable datasets, Zhuang et al. [5] introduced the UAV Distance Estimation (UAVDE) dataset, which utilizes Ultra-Wideband (UWB) technology to acquire distance annotations. This dataset provides a valuable resource for researchers working on UAV target detection and distance estimation. In addition to the dataset, Zhuang et al. also proposed a Positional Correction Module (PCM) to achieve accurate and real-time distance estimation. The PCM aims to overcome the limitations imposed by UAV hardware constraints and enable efficient distance estimation. However, the PCM directly takes the detector's bounding box as input, failing to establish a relationship between object detection and distance estimation. This limitation highlights the need for further research to develop methods that effectively bridge the gap between these two critical tasks in UAV perception. In an object detection pipeline, the initial step involves passing images through a backbone network to generate general features. These features are then fed into different heads, each responsible for predicting classification and bounding box regression results. Since these results often contain redundancies, non maximum suppression (NMS) is employed to filter out redundant boxes. While NMS effectively works for single object detection tasks, our research findings, as depicted in Table 1, indicate that it may not be the optimal choice for distance estimation in the context of UAV targets.

Table1 Analyis study on Bounding Boxes

Model	Val		Test	
	Abs Rel	Sq Rel	Abs Rel	Sq Rel
baseline	0.490	6.716	0.494	6.818
Baseline*	0.448	5.950	0.462	6.220
+PCM	0.148	1.014	0.121	0.620
+PCM*	0.066	0.425	0.058	0.286

Researchers have explored various approaches to improve NMS for object detection task only. These include Greedy-NMS [6] with a fixed threshold that selects the highest-scoring ROI, Soft-NMS [7] and Weighted-NMS [8] that suppress neighboring regions, and AdaptiveNMS [9], which is a dynamic

threshold version of Greedy-NMS. Additionally, DNN based NMS networks have demonstrated superior performance. For example, GossipNet [10] rescores ROIs using their coordinates and scores, Cluster-NMS [11] incorporates geometric information, and Seq2Seq-NMS [12] handles both visual appearance and geometric information of ROIs. The effectiveness of DNN-based NMS networks has been well-established in the literature. Drawing inspiration from these works, we propose a novel approach that leverages feature alignment to utilize the features learned by the detector. Furthermore, we introduce a self-view attention network to capture the relationship between redundant bounding boxes and a cross-view attention network to exploit stereo information. Our method learns to assign appropriate weights to fuse the original bounding boxes, effectively refining the detection results. To achieve accurate distance estimation, we integrate our proposed method with the PCM and train them in an end-to-end manner. This approach enables the learning of optimal bounding boxes tailored specifically for the PCM, thereby enhancing the overall performance of UAV target distance estimation. By establishing a strong connection between object detection and distance estimation through the use of attention mechanisms and feature alignment, our method addresses the limitations of existing approaches and contributes to the advancement of UAV perception.

1 Methods

In this research, our primary objective is to establish a collaborative relationship between an object detector and a stereo distance estimation model. The aim is to enhance the object detector's ability to generate paired matching points that are better suited for accurate stereo distance estimation. To this aim, our methodology comprises several key steps. Firstly, we extract features from the object detector, enabling us to obtain valuable perception information. Subsequently, we introduce two attention mechanisms: a self-view attention mechanism that facilitates the establishment of relationships between different features within a single object, and a multi-view attention mechanism that fosters relationships between stereo images. These attention mechanisms play a critical role in enhancing the object detector's capability to generate matching points suitable for accurate stereo distance estimation. Furthermore, we adopt a comprehensive end-to-end training approach by training the attention modules in conjunction with the distance estimator. This integrated training procedure ensures optimal performance and coherence between the attention mechanisms and the distance estimator

In the subsequent sections, we will provide detailed explanations of each component, elaborating

on their individual roles and contributions within the proposed framework.

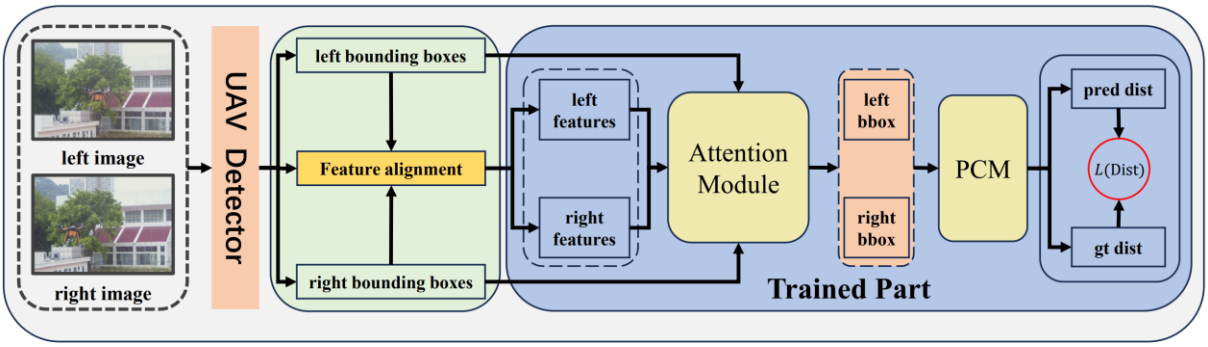

Fig. 1 Entire Pipeline.

1.1 Entire Pipeline

1.1.1 Positional Correction Module

The Positional Correction Module (PCM) is designed to address the position deviation issue in UAV scenarios. This lightweight method can operate on UAV devices in real-time. By taking stereo bounding boxes as inputs, the PCM predicts offsets to compensate for the stereo matching points, enabling accurate distance prediction through triangulation. The PCM is well suited as the distance estimator for our research, as our approach generates reliable bounding boxes that can be used as inputs for the PCM.

1.1.2Positional Correction Module

Initially, we considered training the object detector and distance estimator jointly. However, this approach would require an additional distance label, which is not necessary for the object detector. To improve the system's generality and practicality, we adopted a different strategy: directly extracting features from a pretrained detector.

As illustrated in Figure. 1, our proposed method takes alignment features as input and fuses them with the original detector bounding box to make it suitable for the distance estimator. The parameters of both the proposed method and the PCM are updated using the Mean Squared Error (MSE) Loss during backpropagation. This approach allows for the effective integration of the object detector and distance estimator, while leveraging the pretrained detector's features to enhance the system's performance and adaptability.

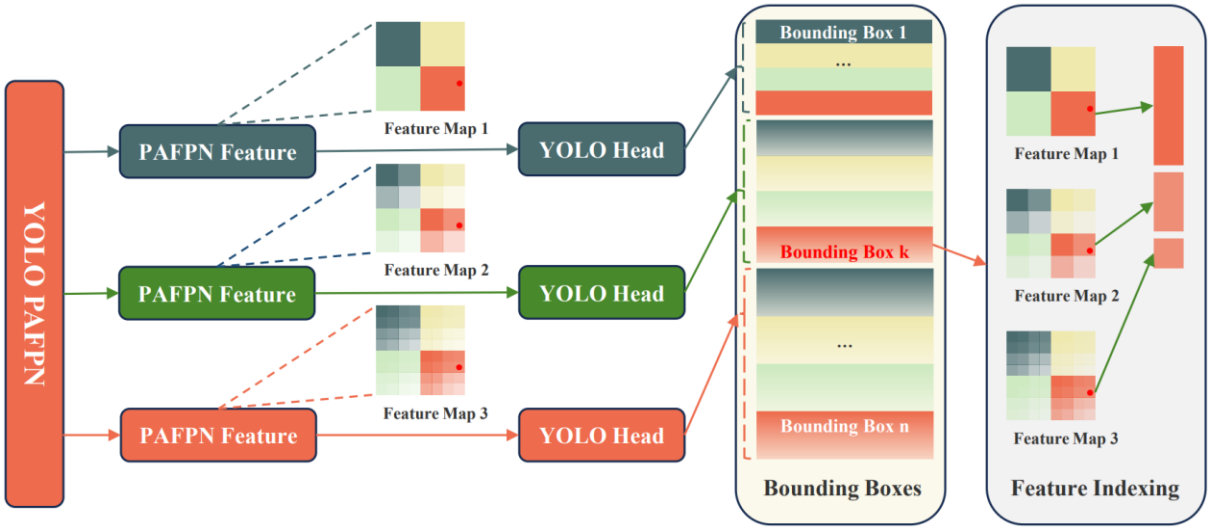

Fig. 2 Illustration of Perception Feature Alignment

1.2 Perception Feature Alignment

Specifically, we select the top 10 boxes for each stereo image, resulting in 10×10 paired points. We compute the distance based on these paired points and observe that the box obtained after applying NMS does not yield the most accurate distance estimation. In contrast to the redundancy-filtering nature of NMS, the features associated with generating these bounding boxes retain distinct information. We refer to the process of extracting features that produce specific box perception as perception feature alignment. The YOLOX architecture comprises three essential components: CSPDarknet serves as the backbone for feature extraction, PAFPN [13] functions as the neck to preprocess the feature and generate three distinct levels of receptive field, and the head is responsible for predicting bounding boxes. Our objective is to align a specific feature computed by PAFPN with its corresponding bounding box. Figure 2 illustrates how a single bounding box corresponds to three different levels of the PAFPN feature, considering the bounding box k marked as red color. The corresponding features to this bounding box are also marked as red dots. Once the features are located, we concatenate them into a single vector. We select the top 10 bounding boxes from both the left and right images, resulting in the alignment of 10 features for both the left and right objects. These features serve as inputs for our proposed attention mechanism.

1.3 Attention Mechanism

The attention mechanism we utilize is specifically designed to capture relationships between sequences. In contrast to NMS, which filters out redundant boxes, we employ our attention mechanism to re-learn the confidence scores of the bounding boxes. These re-learned scores serve as weights for

their corresponding bounding boxes and are then summed to generate a fused bounding box. As shown in Figure 2, we employed self-view attention to learn the representation within each image's features. Once both the left and right features have acquired an intermediate representation, the left intermediate features serve as the queries for the right features. This enables cross-view attention to learn additional representations for re-scoring the right bounding boxes. The right intermediate features undergo a similar process to rescore the left bounding boxes.

2 Exprimental Setting and Result

2.1 UAVDE Dataset

To verify the effectiveness of our proposed method, we utilized the UAV Distance Estimation(UAVDE) dataset [14]. The UAVDE dataset is created for UAV target distance estimation and consists of 3895 stereo images. For the data annotation, the UAVDE dataset applied Ultra Wide Band(UWB) sensors for distance annotation and also manually annotated the UAV bounding boxes on stereo images. The method we propose requires both object bounding box and distance as labels, so the UAVDE dataset is suitable for our research. Specifically, the UAVDE dataset is divided into training, validation, and evaluation datasets, which contain 2815, 541, and 539 stereo images with a resolution of 1280*720 respectively.

2.2 Implementation Details

We adopt YOLOX-Nano as our UAV object detector, following previous work and considering its balance between performance and computational cost, as well as to ensure a fair comparison. We first pre-train the detector on the UAVDE dataset[14], following the same settings as PCM. Our attention mechanism is jointly connected with PCM, enabling end-to-end training and inference. For the training process, we use a batch size of 128 and employ the SGD optimizer with a learning rate of 1e-3, a momentum of 0.9, and a weight decay of 1e-3. We apply gradient clipping with a 1.0 clipping threshold and L2 norm to stabilize the training process. Additionally, we follow the original training protocol, adopting a cosine learning rate schedule with a linear warmup for the first 150 epochs, and train the model for a total of 1500 epochs. For the attention module, we set the inner linear projection dimension to 896, the number of heads to 4, and the dropout to 0.1. Besides, we also incorporate positional encoding to capture spatial information. All experiments are conducted on a single NVIDIA RTX 4090

GPU, enabling efficient training and inference of our model.

2.3 Result and Analysis

We evaluated our method using the Absolute Relative Difference (Abs Rel) and Square Relative Difference (Sq Rel) metrics, consistent with previous studies. To demonstrate the superiority of our approach, we compared it to the state-of-the-art PCM method. Table 2 presents the results, with the baseline being stereo triangulation on UAV detection outcomes. The following observations can be made from these findings: First, the PCM method significantly outperforms the baseline, with an improvement of over 37%, effectively mitigating the issue of position deviation. Second, by incorporating our attention module to generate more accurate bounding boxes for PCM, we achieved a further improvement of 2.1%. This result provides additional evidence supporting the effectiveness and superiority of our proposed approach.

Table1 Reuslts of proposed methods

Model	Val		Test	
	Abs Rel	Sq Rel	Abs Rel	Sq Rel
baseline	0.490	6.716	0.494	6.818
+PCM	0.148	1.014	0.121	0.620
+PCM + Att	**0.116**	**0.547**	**0.101**	**0.309**

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|d^i - d^i_{gt}\right|}{d^i_{gt}} \tag{1}$$

$$\text{SqRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|d^i - d^i_{gt}\right\|^2}{d^i_{gt}} \tag{2}$$

3 Conclusion

In this paper, we addressed the importance of UAV target distance estimation in enabling collective intelligence for swarm UAVs. We proposed a novel approach that leverages self-view and cross-view attention mechanisms to enhance object detection perception, resulting in more accurate

distance estimation. By stacking the attention mechanism and distance estimation model and training them end-to-end, our method effectively learns the relationship between object detection and distance estimation. Experimental results on the UAVDE dataset demonstrate the superiority of our approach, reducing the relative difference from 12.0% to 10.1% compared to the state-of-the-art PCM method. This significant improvement highlights the effectiveness of our proposed approach in advancing drone perception and bringing us closer to realizing the full potential of collective intelligence in swarm UAVs. Future research could explore the application of our approach to real-world scenarios and investigate its scalability and robustness in large-scale swarm UAV systems. Our novel approach contributes to the advancement of UAV target distance estimation and paves the way for the realization of collective intelligence in swarm UAVs.

References

[1] E. Kakaletsis, C. Symeonidis, M. Tzelepi, I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, “Computer vision for autonomous uav flight safety: An overview and a vision-based safe landing pipeline ex ample,” Acm Computing Surveys (Csur), vol. 54, no. 9, pp. 1–37, 2021.

[2] Z. Ma and S. Liu, “A review of 3d recon- struction techniques in civil engineering and their applications,” Advanced Engi- neering Informatics, vol. 37, no. 3, pp. 163–174, 2018.

[3] H. Huang, G. Zhu, Z. Fan, H. Zhai,

Y. Cai, Z. Shi, Z. Dong, and Z. Hao, “Vision-based distributed multi-uav col- lision avoidance via deep reinforce- ment learning for navigation,” in 2022 IEEE/RSJ International Conference on In- telligent Robots and Systems (IROS), pp. 13 745–13 752, 2022.

[4] J. Niemeyer, F. Rottensteiner, and U. So- ergel, “Contextual classification of lidar data and building object detection in urban areas,” ISPRS journal of photogrammetry and remote sensing, vol. 87, pp. 152–165, 2014.

[5] J. Zhuang, D. Yuan, R. Yan, X. Dong, Y. Zhou, W. Huang, and Z. Fan, “Why does stereo triangulation not work in uav distance estimation,” arXiv preprint arXiv:2306.08939, 2023.

[6] P. F. Felzenszwalb, R. B. Girshick,D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2009.

[7] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, “Soft-nms–improving object detection with one

line of code," in Pro- ceedings of the IEEE international confer- ence on computer vision, pp. 5561–5569, 2017.

[8] C. Ning, H. Zhou, Y. Song, and J. Tang, "Inception single shot multibox detector for object detection," in 2017 IEEE In- ternational Conference on Multimedia & Expo Workshops (ICMEW), pp. 549–554, 2017.

[9] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," in Proceedings of the IEEE/CVF conference on computer vision and pat- tern recognition, pp. 6459–6468, 2019.

[10] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4507–4515, 2017.

[11] H. Rezatofighi, N. Tsoi, J. Gwak,

A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regres- sion," in Proceedings of the IEEE/CVF conference on computer vision and pat- tern recognition, pp. 658–666, 2019.

[12] C. Symeonidis, I. Mademlis, I. Pitas, and

N. Nikolaidis, "Neural attention-driven non-maximum suppression for person de- tection," IEEE transactions on image pro- cessing, vol. 32, pp. 2454–2467, 2023.

[13] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8759–8768, 2018.

[14] J. Zhuang, D. Yuan, R. Yan, W. Huang, W. Li, and Z. Fan, "Revisiting stereo triangulation in uav distance estimation," 2023.