DACAM: Depth-Aware Cross-Attention for Defocus-Deblurred Small Object Detection

1st Liang Fan
Department of Computer Science
Loughborough University, UK
L.Fan@lboro.ac.uk

2nd Zhi Chen
School of Computing
Ulster University, UK
Z.Chen@ulster.ac.uk

3rd Baihua Li
Department of Computer Science
Loughborough University, UK
B.Li@lboro.ac.uk

4th Yan Liu
School of Computing and Artificial Intelligence
Southwest Jiaotong University, China
liuyan@my.swjtu.edu.cn

5th Luyang Zhang
zhang.luyang@northeastern.edu

***Abstract*—Object detection in precision agriculture is often severely degraded by defocus blur resulting from unconstrained camera placement and challenging outdoor conditions, which particularly affects small or distant livestock targets, such as calves. To address this, we propose a novel transformer-based detection framework that integrates a geometry-guided vision deblurring transformer network. Our core innovation is a deblurring module that leverages monocular depth estimation to generate geometric priors, which guide a cross-attention mechanism to selectively enhance blurred regions. This targeted feature refinement improves edge and texture clarity in defocused areas, producing detection-friendly features that enable accurate object localization and classification by the subsequent detector. Importantly, our approach allows reliable calf detection using low-cost, fixed-focus cameras in real farm environments, facilitating scalable livestock health monitoring and automated precision agriculture without specialized imaging hardware. We validate our framework on a custom livestock dataset as well as the public DPDD benchmark. Experiments show our integrated model outperforms the RT-DETR baseline by a significant margin, achieving a mean Average Precision (mAP) of 95.7%, particularly with the largest gains observed on small or distant objects, where we observe an improvement of over 10%.**

***Index Terms*—Defocus Deblurring, Depth-Aware Cross-Attention, Targeted Feature Enhancement, Geometry-Guided Vision Transformers.**

I. Introduction

Object detection, a fundamental aspect of computer vision, is essential for applications in precision agriculture, facilitating real-time observation of animal health and behaviour. Nonetheless, real-world applications in settings such as dairy farms frequently encounter uncontrolled conditions and stationary camera positions, resulting in images characterised by considerable defocus blur and low resolution. These degradations present a significant obstacle for contemporary object detectors, as they conceal the intricate features essential for effectively recognising small calf or distant objects. Current object detection frameworks can be classified into two categories: two-stage methods, exemplified by Cascade R-CNN [1], and single-stage methods, represented by the You Only Look Once (YOLO) series [2], [5]. Although efficient, single-stage convolutional models such as YOLO frequently have difficulties with calf object due to continuous downsampling, which reduces spatial resolution and obscures critical edge details. Transformer-based systems such as DETR [3] and its real-time counterpart RT-DETR [7] have demonstrated efficacy in collecting global context. The transformer-based RT-DETR [7] adopted deformable attention to focus on key points in the feature map to enhance detection capabilities for small aerial objects. Nonetheless, their attention systems may exhibit bias towards large, visible things, frequently overlooking the subtler characteristics of smaller targets, particularly when they are obscured. This limitation causes difficulty distinguishing between distant targets and backgrounds, causing the features of objects to be overwhelmed by background noise. Moreover, these models face significant limitations when loading livestock datasets. They are ineffective at detecting small or distant calves, due to defocus and low resolution from livestock monitors. The key features of objects in defocus areas are neglected, resulting in current detection models failing to focus correctly.

To address these limitations, we present an innovative transformer-based deblurring detection model that addresses image blurring within the detection workflow. This model is divided into image deblurring and detection. In image deblurring, we design and train a deblurring module that utilizes the monocular geometry constraints in monocular depth maps as a priori conditions to accurately locate defocus areas. After orientating the defocus region, we utilize the cross attention mechanism to associate the image feature map with the corresponding uncertain depth map. By selectively focusing on information across different regions or channels in the feature space, the cross-attention mechanism achieves inductive bias and local features' representation. In addition, we incorporated the group convolution into the feed-forward network. A group convolution [4] reduces the computational complexity and time required for model processing, while preserving the integrity of the feature map and maintaining the key characteristics of the image. Finally, the deblurred module is added to a detection model. The main contributions of this

work are as follows:

1) We propose a novel deblurring network that leverages monocular depth and associated geometric uncertainty to generate a defocus mask, which guides a cross-attention mechanism to selectively enhance blurred regions. This targeted restoration improves the clarity of edges and textures in defocused areas, enabling more accurate detection of small targets, particularly in farm environments on our livestock dataset by more than 10% mAP, demonstrating its critical role in mitigating defocus blur.
2) We integrate our deblurring network with a transformer-based detector (RT-DETR) in a unified framework. The decoder in the deblurring network further refines image features, enhancing overall clarity before detection. This end-to-end design significantly improves detection performance (from 85.1% to 95.7% mAP), with the largest gains observed in small or distant calves, demonstrating practical benefits for precision livestock monitoring.
3) We introduce a dataset of high-resolution farm images containing calves under diverse environmental conditions. Compared to existing datasets, ours contains 30% more small calves instances, enabling reproducible research in livestock monitoring.
4) Extensive experiments show that the depth-aware deblurring module contributes the largest improvement, with cross-attention and depth uncertainty modeling further refining localization. Ablation results confirm that removing the deblurring module alone reduces the mAP of small rural targets, such as calves, by 10.4%, highlighting its critical role in enhancing the quality of the features for small object detection.

II. RELATED WORK

A. End-to-End Object Detection

End-to-end object detection seamlessly integrates multiple tasks — such as image feature extraction, object classification, and bounding box regression — into a single unified model. This design eliminates the need for separate stages in the detection pipeline, streamlining the process and enhancing efficiency. YOLO [2] revolutionized object detection by introducing a framework that directly predicts the bounding box location and the object category through a single convolution network. The YOLO8 [5] integrates adaptive activation functions and transformer modules. The latest YOLO11 [6] adopts vision transformer (ViT) and hierarchical feature fusion technology. By combining an advanced attention mechanism with the end-to-end detection framework, the YOLO11 achieves higher detection accuracy and speed. The hierarchical feature fusion ensures the information from multiple layers is effectively aggregated to improve the detection of objects at different scales. However, the anchor box in the YOLO series improves the ability of multi-target detection, the size of the non-adaptive anchor box is usually difficult to satisfy the actual size of small targets. The non-adaptive anchor box leads to the error of the target positioning, or the object is missed. The DETR [7] [3] brought a novel structure to strengthen a global feature, including local attention, hierarchical attention, and sparse attention, by introducing the transformer mechanism in CNN-based object detection framework. Moreover, DETR eliminated a non-maximum suppression (NMS) and anchor boxes to simplify the pipeline and enhancing the model's generalization capabilities.

B. Image Deblurring

Image deblurring aims the restoration of a clear image from a blurred input. Through multilayer network architecture on estimating a blur kernel. Recently, the Restormer [11] have attained state-of-the-art outcomes by utilising self-attention to capture long-range dependencies, demonstrating significant efficacy in high-resolution image restoration. Our research expands upon this paradigm by including an innovative steering mechanism derived from geometric clues.

C. Object Detection in Degraded Conditions

Numerous research [8]–[10] has investigated object detection under challenging conditions, including rain, fog, and poor illumination. Nonetheless, defocus blur, particularly in agricultural contexts, remains under investigated. Certain methodologies see restoration and detection as distinct, sequential activities; nevertheless, this may be suboptimal, as the restoration network could fail to retain properties essential for detection. Our approach emphasises a more synergistic integration, wherein the deblurring process is directed by the requirements of the detection task.

III. METHODOLOGY

Our proposed geometry-guided vision deblurring transformer detection framework, illustrated in Figure 1, is composed of three key stages: (1) a deblurring network that uses monocular depth information to guide cross-attention and selectively restore blurred regions. (2) The enhanced features from the deblurring network are further processed in a decoder module, where techniques such as pixel-unshuffle are applied. and (3) a Transformer-based Detector (RT-DETR) performs end-to-end object localization and classification. This pipeline allows for targeted feature restoration in defocused areas, while enabling accurate detection of small or distant objects in a single unified framework.

A. Geometry-Guided Vision Deblurring Transformer

Our geometry-guided vision deblurring transformer employs a transformer-based encoder-decoder architecture aimed at mitigating blur and augmenting attributes essential for detection. The architecture has transformer blocks, each containing two essential components: Depth-Aware Cross-Attention Module and Group-Convolution Feed-forward Network (GCFFN).

1) Depth-Aware Cross-Attention Module: We introduce a depth-aware cross-attention module (DACAM) to address spatially-varying defocus blur in single images. Unlike generic geometry-guided approaches, our module explicitly leverages monocular depth cues to identify defocused regions and

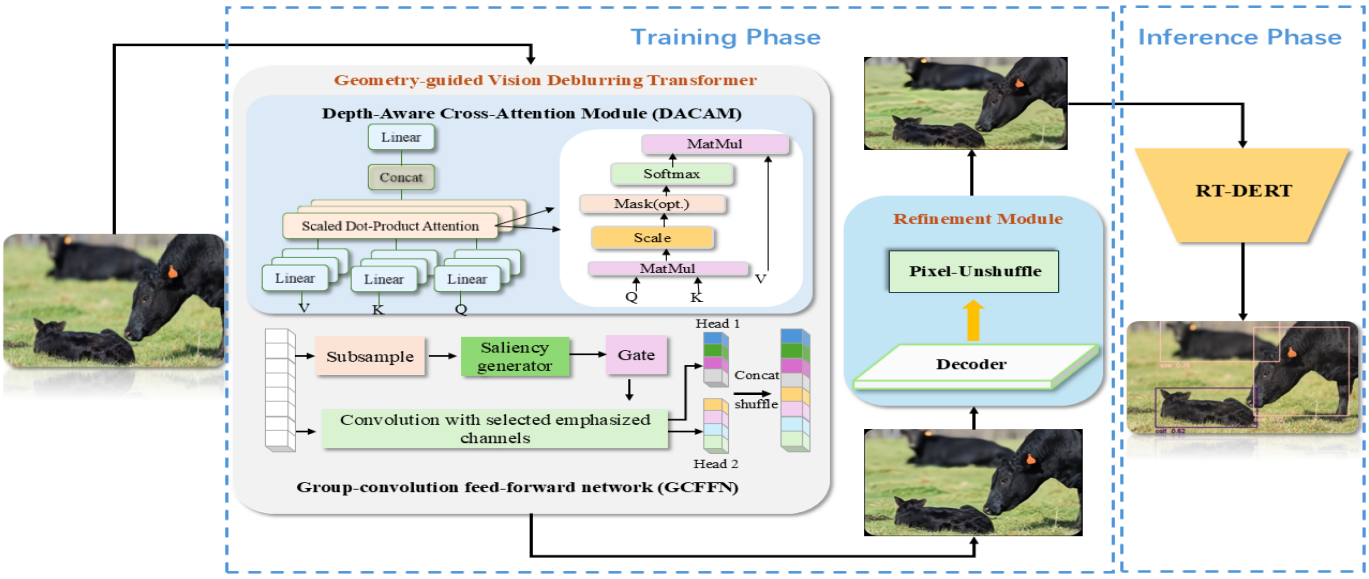

Fig. 1. The pipeline of the depth-aware deblurring detection model. The gray box highlights the deblurring network, which uses monocular depth to derive geometric constraints and guides a cross-attention mechanism to selectively restore blurred regions. The light blue box contains the feature extractor and detector, where the features are further refined to yield clearer object contours and textures, and then passed into a transformer-based backbone (RT-DETR) for object localization and classification. This design allows targeted deblurring while enabling end-to-end object detection.

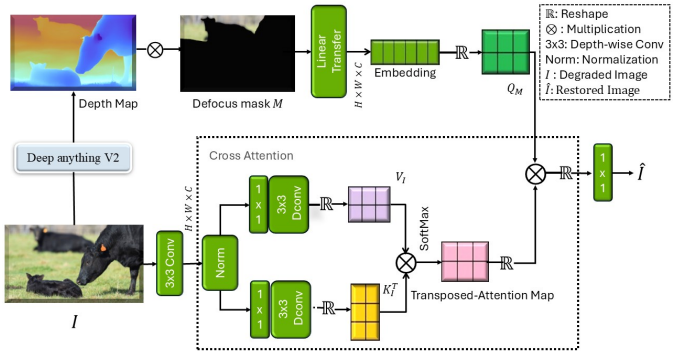

Fig. 2. The structure of Depth-Aware Cross-Attention Module (DACAM). Given an input image, Depth AnythingV2 is first employed to estimate the monocular depth map. Based on the estimated depth, regions that are farther from the camera are identified and used to derive geometric guidance. This guidance is incorporated into a cross-attention mechanism, which selectively focuses on distant defocused areas to refine edge structures and enhance visual clarity. By explicitly leveraging monocular geometry, the module performs targeted sharpening on severely blurred regions.

selectively enhances their features through a cross-attention mechanism. By estimating a single-image depth map, the module identifies depth-dependent blur regions and generates the corresponding defocus mask *M*. This mask steers the cross-attention mechanism to focus restoration specifically on degraded areas. Such a depth-aware strategy not only improves computational efficiency by avoiding global over-processing but also significantly boosts small object calves detection performance by recovering fine structural details based on the scene's underlying 3D geometry. This design ensures that the model adaptively allocates its restorative capacity to out-of-focus areas while preserving the sharp details of in-focus regions. Consequently, this targeted deblurring approach significantly enhances the clarity of calves, providing more reliable features for downstream tasks.

As shown in Figure 2, DACAM deblurring module leverages a cross-attention mechanism to selectively enhance features in defocused regions. The key idea is to exploit geometric cues from monocular depth to identify areas with high blur or depth discontinuities, which are typically challenging for small or distant calves detection. To achieve this, we first estimate a depth map D from the input image I using a pre-trained DepthAnythingV2 model [13]. The depth map serves as a geometric prior, allowing the module to identify depth uncertainties and sharp transition gradients. Formally, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the depth map D which is calculated from Depth AnythingV2 is:

$$D = f_{\text{depth}}(I; \theta_{\text{pretrained}}), \tag{1}$$

where f_{depth} denotes the pre-trained depth estimation network with parameters $\theta_{\text{pretrained}}$.

Next, a defocus mask M is generated from the depth map to act as a spatial attention proxy. Specifically, we compute the gradient magnitude of the depth maps:

$$\nabla D = \|(D_x, D_y)\|_2, \tag{2}$$

where (D_x, D_y) are depth derivatives along the horizontal and vertical directions. High gradient values correspond to depth discontinuities or out-of-focus regions, indicating areas where the point spread function (PSF) has attenuated blurred boundaries.

We take advantage of the principle that depth estimation is often uncertain and produces high gradients in out-of-focus regions. The binary defocus mask M is then generated by thresholding the depth gradients, where regions with gradients exceeding a threshold τ are considered defocused:

$$I_{\text{Grad}}(i,j) = \mathbf{1}_{\{\nabla D(i,j) < \tau\}} \tag{3}$$

To explicitly address spatially-varying defocus blur, we employ a cross-attention mechanism guided by the defocus mask M. The purpose of this attention is to selectively concentrate the network's restorative capacity on physically blurred regions, rather than applying uniform enhancement across the entire feature map. By using the defocus mask derived from monocular depth, the query Q_M focuses on degraded areas, while the key K_I and value V_I capture image features. The cross-attention operation:

$$\text{CrossAttention}(Q_M, K_I, V_I) = \text{Softmax}\Big(\frac{Q_M K_I^\top}{\sqrt{d_k}}\Big) V_I \tag{4}$$

where d^k is the dimension of the key K_I. The cross-attention ensures that the network enhances edges and textures where blur is severe, preserves in-focus regions, and produces high-quality, detection-friendly features. This design is particularly beneficial for small or distant livestock detection, where recovering fine structural details is critical.

By doing so, DACAM selectively concentrates its deblurring capability on the physically blurred regions, while leaving in-focus areas largely untouched. This targeted feature enhancement recovers sharp contours and texture details in defocused regions, producing high-quality, detection-friendly features that significantly improve small calves detection performance. In summary, our DACAM module differs from generic geometry-guided designs by leveraging monocular depth to accurately localize defocus blur, guiding the restoration process precisely to the degraded regions. Through a cross-attention mechanism, the network adaptively enhances these blurred areas rather than applying uniform sharpening, producing enriched features that preserve structural details and sharp contours. These high-quality, detection-friendly features are particularly critical for accurate small calves detection in challenging, out-of-focus regions.

2) Group-Convolution Feed-forward Network (GCFFN): To enhance efficiency and preserve local feature structure, we replace the standard feed-forward network in the transformer block with a Group-Convolution Feedforward Network (GCFFN). Conventional transformer FFNs rely on fully connected layers, which are computationally expensive for high-resolution feature maps and do not explicitly capture local spatial structure, making them inefficient for real-time or edge-deployed vision systems. In contrast, GCFFN integrates CNN-style locality with transformer attention: The input feature maps are divided into several groups, and convolutions [4] are applied separately to each group before concatenating the results. The final output Y from the group convolution strategy is the concatenation of the convolution within g groups.

$$Y = [Y_1, Y_2, \ldots, Y_g] \tag{5}$$

The output of convolution on the g channel is $Y_g = \text{Conv}(\hat{x}_g, W_g)$. This grouped convolution strategy preserves fine textures and edge details critical for deblurring, while significantly reducing computational cost compared with fully connected layers. By maintaining rich local structure within high-resolution feature maps, GCFFN produces detection-friendly features that are particularly beneficial for accurate calves detection in agricultural images, such as calves and distant livestock, where fine structural details are essential.

B. Refinement Module and Detector

The enhanced feature maps produced by our deblurring network are passed into a decoder module, which further refines the image. Inspired by high-resolution image generation techniques commonly used in GANs, we apply a pixel-unshuffle operation [14] which rearranges feature map channels into a higher spatial resolution while preserving high-frequency information, to progressively enhance spatial details across the entire image, effectively sharpening edges and textures. This design allows the network to recover fine-grained structures and high-frequency details that are critical for detecting calves, distant livestock targets, such as calves, in defocused regions. The RT-DETR detector utilises a transformer encoder-decoder to process these data, predicting bounding boxes and class labels for all objects within the image.

IV. EXPERIMENTS AND RESULTS

A. Datasets and Implementation Details

Livestock Dataset: The images of livestock dataset was produced from 14 multi-view cameras at a dairy farm. All cameras capture images of cows and calves under natural farming conditions. Due to the structural constraints of the facility, the cameras are positioned at the corners and along the roof edges, resulting in animals often appearing at a considerable distance from the lens or in partially focused regions. Additionally, animals that are partially obscured by obstacles such as trees, fences, or larger animals pose a significant challenge in this dataset. Furthermore, the close interactions between animals in the pasture, such as moving together or crossing paths, further complicate identification. We selected 66 representative video clips from all cameras that contain scenes of both adult cows and newborn calves. To construct the dataset, we extract key frames from these clips and removed redundant frames using a frame-difference filter. To further reduce data redundancy, we retain only those with significant differences using the filter on all extracted frames, if the information from frames is similar. After filtering, the number of our livestock images comprise 1,939 training images, 488 validation images, and 102 testing images of cows and calves in demanding, lifelike situations. The dataset exhibits considerable class imbalance (85% cows, 15% calves) along with recurrent defocus blur and occlusion.

The dataset will be made publically available upon acceptance.

DPDD Dataset: To benchmark our method on a public dataset, we employ the DPDD dataset [16], which exhibits real-world defocus blur and resolutions comparable to our

TABLE I

COMPARISON WITH DEBLURRING MODULE OF DETECTOR FOR DIFFERENT OBJECTS (INCLUDING MULTI SIZE AND SMALL SIZE CALVES) ON LIVESTOCK DATASET

Models	mAP_{val50}		$mAP_{val50:95}$	
	Multi	***Small calves***	***Multi***	***Small calves***
YOLO11 [6]	0.605	0.519	0.270	0.226
YOLO11+SAHI	0.855	0.795	0.445	0.399
RT-DETR [3]	0.858	0.792	0.483	0.407
RT-DETR+SAHI	0.920	0.834	0.575	0.500
Ours	0.957	0.896	0.618	0.543

images. The DPDD dataset comprises 350 images that are utilised for training, while the remaining 150 images are utilised for validation and testing, respectively. These images in the DPDD dataset cover diverse indoor and outdoor scenes with spatially varying blur, posing challenges for restoring fine features, which is critical for accurate downstream detection.

Livestock Dataset and DPDD Dataset provide both realistic agricultural scenarios and well-established defocus benchmarks, allowing us to evaluate the effectiveness of our geometry-guided deblurring and detection framework in both domain-specific and general settings.

Implementation Details: The model undergoes training for 90 epochs via the AdamW optimiser [15], with a batch size of 8 and an initial learning rate of 0.001.

B. Quantitative Results

We compare our model with several baselines on our livestock dataset, including YOLOv11 and RT-DETR. We also report results with Slice-Assisted Hyper Inference (SAHI) [12], for improving calves detection. Table 1 demonstrates that our technique outperforms all baseline models. Significantly, for diminutive items (calves), our model attains a mean Average Precision (mAP) of 54.3%, illustrating the efficacy of geometry-guided deblurring in retrieving characteristics crucial for identifying difficult objects. We also compare our model with other baselines on the DPDD dataset. Table 2 demonstrates that our deblurring module achieves a PSNR of 28.93, which is comparable to the leading Restormer model, hence confirming the efficacy of our suggested architecture for deblurring. A smaller SSIM facilitates superior preservation of image structure to enhance the robustness. Figures 3 (a) and 3 (b) illustrate the efficacy of our method in recovering essential features and enhancing model performance under demanding conditions. The photos produced by our method are not only visually appealing but also demonstrate improved global and local features, yielding outputs that seem more realistic than those generated by alternative techniques. This underscores the capacity of our deblurring technique to harmonise intricate details with overall visual consistency.

Figure 4 illustrates the assessment performed using our transformer-based detection model. We utilised the pre-trained transformer-based deblurring model on our dataset to mitigate the effects of defocus blur. The increase of 1. 65% in PSNR signifies a focused restoration of critical image areas impacted by defocus, thus facilitating enhanced feature extraction and detection. Restored critical image regions allows the downstream transformer detector to extract more discriminative features, leading to higher localization and classification accuracy, particularly for small or distant calves. This demonstrates that our deblurring module not only improves visual quality but directly contributes to robust and accurate calves detection in challenging livestock environments.

TABLE II

THE QUALITATIVE EVALUATION ON THE DPDD DATASET. THIS TABLE COMPARES OUR METHOD WITH SEVERAL STATE-OF-THE-ART DEBLURRING METHODS ON THE DPDD DATASET USING PSNR, SSIM, LPIPS, AND MAE AS EVALUATION METRICS.

Methods	***PSNR***	***SSIM***	***LPIPS***	***MAE***
Restormer [11]	28.87	0.882	0.145	0.025
DPDNetD [16]	24.34	0.747	0.277	0.044
RDPDNet+ [17]	25.39	0.772	0.319	0.040
Ours	28.93	0.859	0.224	0.027

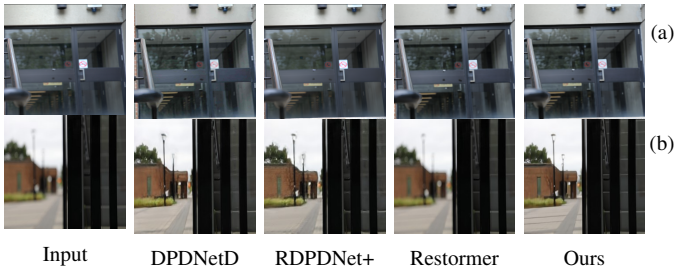

Fig. 3. The qualitative comparison on the DPDD dataset. The first column shows the defocused input images, and the last column presents the results of our method. The remaining columns display deblurring results produced by different competing methods. (a) An example of long-distance defocus blur, where the text on the distant flag is severely blurred. Our method effectively restores the fine details, making the characters on the flag significantly clearer. (b) An example of long-distance defocus blur on buildings, where the contours of distant houses are indistinct. Our method recovers sharper structural boundaries and clearer building outlines compared with other methods.

Fig. 4. Detection performance on our livestock dataset with and without our deblurring network. The bottom-right corner of each image shows a zoomed-in view of targets located in the distant defocused regions, which facilitates clearer visual inspection. (a) Detection results without applying our deblurring method, where targets in distant defocused areas are prone to missed detections. (b) Detection results with our deblurring method, which enhances the clarity of distant defocused regions and effectively alleviates missed detections of small or blurred targets.

C. Ablation Studies of our deblurring module

In the ablation experiment as shown in Table 3, we utilised a convolutional U-Net [18] as a baseline to assess the effi-

TABLE III
ABLATION STUDIES OF DEBLURRING MODULE USING PSNR.

Networks	***Components***	***PSNR***
Baseline	U-net based on Convolution	24.37
Attention	Cross attention+ FFN	26.26
Improved Attention	DACAM+ FFN	27.91
Feed-forward Network	DACAM + linear GCFFN	28.79
The Overall of our method	DACAM + nonlinear GCFFN	28.93

TABLE IV
ABLATION STUDIES OF DETECTION ACCURACY

Networks	***Deblurring Strategy***	mAP_{val50} ***(overall)***	mAP_{val50} ***(Small calves)***
Baseline(RT-DETR)	None	0.858	0.792
Baseline+UNet	UNet	0.860	0.802
Baseline+Attention	Cross attention+ FFN	0.893	0.856
Baseline+Improved Attention	DACAM + FFN	0.926	0.858
Baseline+Feed-forward Network	DACAM +linear GCFFN	0.946	0.890
Baseline+Ours	Depth-Aware Deblurring	0.957	0.896

cacy of our proposed deblurring network. We systematically substituted the convolutional blocks in the U-Net with our attention modules to evaluate the effects of our alterations. The integration of the cross-attention module resulted in a PSNR enhancement to 26.26, without the guidance of geometry constraints. Moreover, the enhancement of the cross-attention module with geometry constraints-guided information resulted in a further increase in PSNR of 27.91. The PSNR value underscores the significance of utilising depth information for enhanced accuracy in defocus deblurring. To enhance performance, we substituted a feed-forward neural network with a nonlinear group convolution module, resulting in a PSNR rise to 28.93. We further compare our proposed depth-aware deblurring detection framework with conventional "deblur-then-detect" pipelines. As shown in Table 4, applying a generic deblurring network prior to RT-DETR improves PSNR and overall mAP, but the improvement on small or distant calves object remains limited. In contrast, our method, which integrates depth-guided geometric priors into a cross-attention mechanism, selectively enhances blurred regions. This targeted feature refinement leads to substantial gains in small calves detection, demonstrating the unique advantage of geometry-guided attention in leveraging spatial structure for robust and accurate object detection.

V. CONCLUSION

This paper presents a geometry-guided vision deblurring transformer detection model for blurred images. Initially, monocular geometric restrictions derived from depth maps are employed within a cross-attention mechanism to mitigate blur in defocused images. The improved features are derived from deblurred photos. A specialised deblurring detection network is trained with the DERT detector to accommodate cattle data. Our suggested detection methodology enhanced detection accuracy to 95.7% in our cattle dataset when merged with the RT-DETR-based SAHI inference model. The detection accuracy for calves as objects improved to 89.6% using our model, compared to 79.2% with the RT-DERT model. This demonstrates the efficacy of our method in enhancing detection reliability inside intricate and dynamic situations. While currently validated on livestock scenarios, generalization to other blur types or environments remains to be explored. Future work will focus on real-time deployment with a lightweight transformer using single-view geometric priors. The method could be embedded into existing livestock monitoring systems to enable automated calf health tracking, early anomaly detection, and reduced manual inspection, enhancing both operational efficiency and animal welfare.

REFERENCES

[1] Zhaowei Cai et al.: Cascade R-CNN: Delving into High Quality Object Detection. In: 2010 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6154-6162. IEEE, (2010)

[2] Redmon, J.: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788. IEEE, (2016)

[3] Yian Zhao et al.: DETRs Beat YOLOs on Real-Time Object Detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965-16974. IEEE, (2024)

[4] Zhuo Su et al.: Dynamic Group Convolution for Accelerating Convolutional Neural Networks. In: European Conference on Computer Vision, pp. 138-155. Springer, (2020)

[5] Rejin Varghese et al.: YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), pp. 1-6. IEEE, (2024)

[6] Khanam, Rahima and Hussain, Muhammad: YOLOv11: An Overview of the Key Architectural Enhancements. arXiv preprint arXiv:2410.17725. (2024)

[7] Khanam, Rahima and Hussain, Muhammad: Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv preprint arXiv:2010.04159. (2020)

[8] Thomas Eboli et al.: End-to-End Interpretable Learning of Non-Blind Image Deblurring. In: European Conference on Computer Vision, pp. 314-331. Springer, (2020)

[9] Yanni Zhang et al.:Multi-Scale Frequency Separation Network for Image Deblurring. IEEE Transactions on Circuits and Systems for Video Technology **33**(10), 5525–5537 (2023)

[10] Dongxiao Zhang et al.:Joint Motion Deblurring and Super-Resolution for Single Image Using Diffusion Model and GAN. IEEE Signal Processing Letters **31**, 736–740 (2024)

[11] Syed Waqas Zamir et al.: Restormer: Efficient transformer for high-resolution image restoration. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5728-5739. IEEE, (2022)

[12] Chollet, François: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251-1258. IEEE, (2017)

[13] Lihe Yang et al.: Depth Anything V2. arXiv preprint arXiv:2406.09414. (2024)

[14] Bin Sun et al..: Hybrid pixel-unshuffled network for lightweight image super-resolution. In: 2023 The Association for the Advancement of Artificial Intelligence (AAAI), pp. 2375-2383. (2023)

[15] Loshchilov, I.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. (2017)

[16] Abuolaim, Abdullah and Brown, Michael S.: Defocus deblurring using dual-pixel data. In: European Conference on Computer Vision, pp. 111-126. Springer, (2020)

[17] Abdullah Abuolaim et al.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: 2021 IEEE/CVF Conference on Computer Vision, pp. 2289-2298. IEEE, (2021)

[18] Olaf Ronneberger et al.:U-net: Convolutional networks for biomedical image segmentation. In: 2015 International Conference on Medical image computing and computer-assisted intervention (MICCAI), pp. 234-241. Springer, (2015)