bioengineering

MDPI

Article

GRIT-OBB: Geometry-Guided Oriented Object Detection for Fetal Cardiac Structures in Ultrasound

Jiachen Yuan [1], Wenzheng Huang [1], Anteng Cai [1], Qiujin Wang [1], Zihan Ye [1], Ning Xin [1], Camilla Fazi [2,3], Federica Fontanella [4] and Netzahualcoyotl Hernandez-Cruz [1,*]

[1] Academy of AI and Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215000, China; jiachen.yuan23@student.xjtlu.edu.cn (J.Y.); wenzheng.huang23@student.xjtlu.edu.cn (W.H.); anteng.cai23@student.xjtlu.edu.cn (A.C.); qiujin.wang23@student.xjtlu.edu.cn (Q.W.); zihan.ye23@student.xjtlu.edu.cn (Z.Y.); ning.xin21@student.xjtlu.edu.cn (N.X.)
[2] Department of Neurosciences, Psychology, Drug Research and Child Health, University of Florence, 50139 Florence, Italy; camilla.fazi@unifi.it
[3] Division of Neonatology, Careggi University Hospital of Florence, 50134 Florence, Italy
[4] Department of Obstetrics and Gynecology, University Medical Center Groningen, 9713 GZ Groningen, The Netherlands; f.fontanella@umcg.nl
* Correspondence: n.hernandezcruz@xjtlu.edu.cn

Abstract

Fetal cardiac structure localization requires features that follow the size and orientation of individual anatomical targets. We present Geometry-guided Regional Interaction and Transformation for Oriented Bounding Box Detection (GRIT-OBB), an oriented method that uses corrected proposal geometry to guide feature extraction and bounding-box refinement. Starting from horizontal proposals, a supervised correction stage estimates oriented bounding boxes, which define the sampling regions for rotated position-sensitive alignment. Then, a geometry-guided relation residual combines regional features using Gaussian bounding box representations and Wasserstein distances, followed by classification and refinement. Finally, rotated intersection-over-union losses directly supervise both proposal correction and localization. We evaluated GRIT-OBB against six methods on three datasets: FOCUS, OCD, and AU-OPG, comprising 300, 120, and 901 images, respectively. On the FOCUS dataset, GRIT-OBB achieved a mean AP50 of 47.74%, AP75 of 8.53%, and mAP50:95 of 16.81%, corresponding to improvements of 4.74%, 1.8%, and 2.01% over the strongest baseline, respectively. On the OCD dataset, GRIT-OBB achieved a mean AP50 of 73.14%, AP75 of 18.97%, and mAP50:95 of 30.77%, corresponding to improvements of 12.40%, 13.54%, and 11.28% over the strongest baseline. On the AU-OPG dataset, GRIT-OBB achieved a mean AP50 of 82.65%, AP75 of 59.66%, and mAP50:95 of 50.66%, corresponding to an improvement of 2.41%, 7.27%, and 3.88% compared with the strongest baseline. In an ablation study, removing supervised geometric correction reduced performance on the FOCUS dataset from 47.74% to 43.44% for AP50, while removing both IoU losses decreased performance from 8.53% to 6.59% for AP75. These results support our hypothesis that geometry-guided feature alignment and direct supervision at successive localization stages can improve object detection for fetal cardiac structures.

Keywords: fetal cardiac ultrasound; oriented object detection; anatomical localization; feature alignment; geometric relations; rotated intersection over union

Check for updates

Academic Editor: Alfonso Maria Ponsiglione

Received: 29 September 2026
Revised: 26 October 2026
Accepted: 29 October 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Accurate localization of fetal cardiac structures provides the anatomical reference needed for structure-specific assessment and measurement. Detecting multiple structures within the same ultrasound image requires features that distinguish each target while preserving its relationship to the surrounding anatomy. Research in fetal echocardiography has explored attention mechanisms, coarse-to-fine processing, and spatiotemporal modeling for anatomical classification and segmentation [1,2]. These developments place anatomical representation at the center of computational ultrasound analysis, where models distinguish individual structures, extract features from their spatial support, and retain the regional information needed to interpret targets within the surrounding anatomy.

Recent work has also combined geometric state estimation with multimodal interpretation in sensor-equipped systems. For example, Mao et al. integrated flexible sensing and shape reconstruction with a fine-tuned multimodal large language model for state interpretation and fault diagnosis in a six-bar tensegrity system [3]. In fetal ultrasound, the spatial representation must capture the size, orientation, and relationships of anatomical structures within the image.

For object detection, cardiac structures pose several challenges: they differ in size and shape; several of these targets are nested, adjacent structures share image context, and changes in orientation alter the tissue enclosed by a rectangular region. Low contrast and similar local appearances further challenge the recognition of small targets [4–7]. Related work has used anatomical localization to identify relevant content in fetal ultrasound videos and support cardiac measurements. TIER-LOC uses visual queries to locate relevant temporal intervals [8], while automated cardiac function analysis combines structure detection and segmentation with geometric measurement [9]. Our task concerns the spatial localization of multiple anatomical structures within a single image. Each structure must be distinguished from neighboring anatomy and localized according to its size and orientation. An oriented bounding box (OBB), parameterized by its center, side lengths, and angle, provides a compact representation of these properties.

Using OBBs for fetal cardiac structure detection requires more than predicting an angle at the final output. Small and oblique structures require features sampled from regions that match their spatial extent, while their similar local appearances make contextual information from neighboring regions useful. When an intermediate bounding box defines the sampling region, its geometric accuracy also affects the features available for subsequent prediction. We therefore identify three requirements: correspondence between target geometry and regional feature extraction [10–12], spatially informed interaction among regions [13,14], and direct supervision of intermediate and final localization [15,16].

The first requirement is that the target geometry corresponds to the regional feature extraction. A horizontal proposal enclosing an oblique structure can include substantial surrounding tissue, making the sampled representation sensitive to the proposal's orientation and extent. RoI Transformer addresses this mismatch through supervised conversion from horizontal proposals to oriented regions [10]. R3Det and S^2A-Net further develop feature refinement and alignment mechanisms for oriented detection [11,12]. For fetal cardiac structures, this correspondence is particularly important when a small or elongated structure shares its image region with neighboring anatomy. Aligning feature extraction with the target's orientation and extent yields a sampling frame that more faithfully reflects the structure being localized.

The second requirement is spatially informed interaction among regions. A small target may provide limited local evidence, while neighboring regions contain information useful for distinguishing its anatomical class. Relation Networks model interactions between objects using both appearance features and geometry [13]. For oriented objects,

OrientedFormer introduces Gaussian positional encoding and Wasserstein-based attention to represent bounding-box geometry and condition interactions between queries [14]. These developments motivate the use of corrected proposal geometry to organize contextual feature exchange after target-aligned sampling.

The third requirement is direct supervision of successive geometric estimates. An intermediate bounding box determines where the next stage extracts features, whereas the final prediction determines the reported localization. Cascade R-CNN trains successive detection stages with increasing Intersection over Union (IoU) thresholds, demonstrating the importance of matching stage-specific training to the quality of incoming proposals [15]. GIoU addresses the mismatch between coordinate-based regression losses and overlap-based evaluation by introducing an overlap-based regression objective [16]. These findings motivate direct localization supervision at both stages: the corrected proposal should overlap its assigned ground truth to provide an appropriate sampling reference, and the final prediction should be supervised against its assigned ground truth to improve the reported localization.

We address these three requirements using GRIT-OBB, a method for detecting oriented fetal cardiac structures. A ResNet-50 backbone and feature pyramid generate features for horizontal region proposals. A supervised geometry learner converts these proposals into oriented references and rotated position-sensitive alignment, then samples features in their local coordinate frames. Building on Gaussian geometric representations, the geometry-guided relation residual uses positional encodings and a Wasserstein-based attention bias to exchange information among the aligned regions. A learned gate controls the contextual update to each regional feature. Parallel classification and bounding-box refinement heads then identify anatomical classes and refine their geometry. Equivalent regression and direct-rotated intersection-over-union losses supervise both the corrected proposals and the final detections. The corrected references are detached before subsequent sampling, and each stage receives its own direct localization supervision. The proposed method uses learned proposal geometry to organize where features are extracted, how regions interact, and how localization is refined.

GRIT-OBB builds on established oriented proposal transformation and feature alignment [10] and Gaussian positional encoding and Wasserstein-based attention [14]. We do not claim these techniques or IoU-based supervision individually as new. Our contributions concern how corrected proposal geometry connects feature extraction, regional interaction, and localization supervision:

1. A geometry-guided detection method in which the same corrected oriented proposals provide the spatial reference for rotated position-sensitive alignment, regional interaction, and final bounding-box refinement. This organization links the geometry used to extract regional features with the geometry used to condition contextual interaction and refine fetal cardiac structure localization.
2. A geometry-guided relation residual that operates after rotated feature alignment and conditions interactions among the extracted regional features on corrected proposal geometry. It combines Gaussian positional encoding and a Wasserstein-based attention bias with a learned gate that controls the contextual residual added to each aligned feature.
3. A two-stage localization supervision scheme tailored to this geometry-guided pipeline. Equivalent bounding-box regression and direct rotated IoU supervision both train the corrected proposals and the final predictions against their respective assigned ground truth. This gives the intermediate sampling reference its own localization objective despite its detached connection to subsequent feature alignment.

2. Related Work

We first review anatomical localization in fetal cardiac ultrasound, then examine three technical foundations of GRIT-OBB: oriented detection and feature alignment, geometric representation and regional interaction, and localization objectives and geometric supervision.

2.1. Anatomical Localization in Fetal Cardiac Ultrasound

An ensemble approach combines cardiac view recognition with prenatal detection of complex congenital heart disease [17], while automated cardiac function analysis integrates structural detection and segmentation with geometric measurements [9]. These applications connect anatomical image analysis with diagnostic assessment and quantitative measurement. However, the segmentation evaluation in the latter paper covers the four cardiac chambers, the heart, and the thorax, but does not assess segmentation performance for valves, septa, or small anatomical landmarks [9]. Localizing these additional targets requires distinguishing structures with different spatial extents and similar appearances within the same cardiac image.

For example, TIER-LOC uses visual queries to identify temporal intervals containing relevant anatomical content in fetal ultrasound videos [8]. While temporal localization identifies when the requested anatomy appears, it does not specify the oriented spatial extent of individual structures within a frame. In this paper, we address the spatial localization of multiple fetal cardiac structures within individual ultrasound images, assigning each detected region an anatomical class and an oriented bounding box. GRIT-OBB uses corrected proposals to align feature extraction with individual targets and to incorporate geometric regional interactions and contextual evidence from surrounding anatomy.

2.2. Oriented Detection and Feature Alignment

Detection architectures differ in how they organize candidate locations and construct features for prediction. For example, RetinaNet uses focal loss to address foreground–background imbalance in dense detection [18], while FCOS predicts objects through a fully convolutional, anchor-free formulation [19]. Region-based approaches construct a representation for each proposal before classification and refinement. However, horizontal sampling regions can include substantial background around oblique objects, creating a mismatch between the target geometry and the extracted features [10,12].

RoI Transformer learns the conversion from horizontal proposals to oriented regions and uses the transformed geometry for regional feature extraction [10]. For example, Oriented R-CNN generates oriented proposals within its region proposal network and refines them through an oriented detection head [20]. These approaches introduce orientation at different stages: RoI Transformer converts horizontal proposals into oriented regions after proposal generation, whereas Oriented R-CNN generates oriented proposals within the RPN. R3Det couples successive bounding-box refinement with feature refinement [11], while S^2A-Net explicitly aligns features with oriented anchors [12]. Together, these methods establish the importance of matching feature extraction to the geometry used for prediction.

Transformer-based detectors offer additional ways to represent oriented targets. For example, Oriented-DETR uses a point-axis representation to encode oriented objects [21]. OrientedFormer combines Gaussian positional encoding, geometric self-attention, and oriented cross-attention to connect object geometry with query interaction and feature sampling [14]. O2-RT-DETR develops a real-time transformer formulation for oriented detection in remote sensing images [22]. However, OrientedFormer reports remaining difficulties with small objects, extreme aspect ratios, and foreground objects that resemble the background [14]. These cases are relevant to fetal cardiac localization, in which small or

elongated structures share imaging context with neighboring anatomy. GRIT-OBB combines sampling in corrected proposal frames with geometric interaction among regional features to incorporate both target-aligned detail and surrounding context.

2.3. Geometric Representation and Regional Interaction

The orientation of a bounding box affects how geometric differences are measured. For example, Gaussian Wasserstein distance (GWD) loss maps rotated rectangles to two-dimensional Gaussian distributions and measures the discrepancy between them in distribution space, thereby addressing discontinuities associated with oriented bounding-box parameterizations [23]. GauCho parameterizes Gaussian distributions through Cholesky decomposition for oriented object detection [24]. While these representations encode position, extent, and orientation, a geometric distance alone does not determine how appearance information should be exchanged among candidate regions.

Context can guide feature modeling through semantic modulation or structured knowledge injection [25,26]. Another example is Relation Networks, which combine appearance and geometric information when exchanging features among object candidates [13]. For example, OrientedFormer extends this approach to oriented queries through Gaussian positional encoding and Wasserstein-based attention [14]. Its decoder performs geometric self-attention among queries before oriented cross-attention samples image features. These operations connect regional geometry with visual evidence; however, their organization differs from that of interactions among features already extracted within corrected proposal frames.

In contrast, our method (GRIT-OBB), first applies rotated position-sensitive alignment to corrected oriented proposals, then exchanges information among the resulting regional features. The corrected proposals supply both the sampling frames and the Gaussian geometry used to condition attention. A learned gate controls the contextual residual added to each aligned feature. This design places regional interaction after explicit target-aligned sampling and uses a shared geometric reference for both operations.

2.4. Localization Objectives and Geometric Supervision

Bounding-box regression must link its optimization objective to the overlap measure used to evaluate localization. For example, Generalized IoU addresses the discrepancy between coordinate-based regression losses and overlap quality through an objective that also accounts for non-overlapping axis-aligned bounding boxes [16]. Rotated detection introduces further difficulties because different combinations of side lengths and angles can describe equivalent rectangles. GWD addresses parameterization discontinuities through a distributional regression objective [23]. However, this objective measures distance between Gaussian representations rather than directly computing the intersection and union of the corresponding rectangles.

The form of the regression loss and the supervision of successive estimates address different aspects of localization. For example, Cascade R-CNN trains successive detection stages with increasing IoU thresholds to match stage-specific training to the quality of incoming proposals [15]. While this approach emphasizes progressively refined detections, a geometry-guided sampling stage also requires an accurate intermediate reference because its coordinates determine where subsequent features are extracted.

GRIT-OBB combines equivalent bounding-box regression with direct rotated IoU losses at the proposal-correction and final-refinement stages. Equivalent bounding-box regression accommodates alternative parameterizations of the same rectangle, while the overlap terms supervise each geometric estimate against its assigned ground truth. The correction objective trains the reference used for feature sampling, and the final objective

trains the reported detection. Both stages receive direct localization supervision, with the corrected reference detached before subsequent sampling.

These reviewed papers motivated our hypothesis that using supervised proposal geometry to guide regional feature extraction and contextual interaction, together with direct supervision of intermediate and final predictions, can improve fetal cardiac structure localization. To investigate our hypothesis, we developed GRIT-OBB, a geometry-guided oriented detection method that uses corrected proposals as a shared spatial reference for rotated feature alignment, geometric regional interaction, and final bounding-box refinement, and applies direct rotated IoU supervision to both the corrected proposals and final predictions.

Section 3 presents the datasets, preprocessing, and proposed methodology. Section 4 introduces the experimental settings and evaluates detection performance, robustness, and component contributions. Section 5 discusses the findings and limitations, and Section 6 concludes the paper.

3. Methodology

3.1. Data and Preprocessing

3.1.1. Datasets and Fixed Partitions

We used three open-access datasets: FOCUS [27], consisting of fetal ultrasound images. OCD microscopy [28], consisting of microscopy images and AU-OPG panoramic [29,30], consisting of dental radiograph images.

We used the open-access FOCUS [27] dataset, comprising 300 images from 217 participants aged over 18 years, collected during the second trimester at two medical centers. No new data were collected. The dataset was released with a training/validation/test partition of 200/50/50 images performed at the participant level. The FOCUS images are PNG files with resolutions ranging from 640 × 392 to 1137 × 787 pixels (mean: 956 × 655, SD: 88 × 74) and include segmentation masks for the thorax and heart. To support the objectives of this paper, we enriched the dataset with bounding boxes for 14 anatomical structures, in addition to the whole heart: atrioventricular valves offset (AVO), cardiac apex (Apex), descending aorta (DA), foramen ovale (FO), interatrial septum (IAS), interventricular septum (IVS), left atrium (LA), left ventricle (LV), mitral valve (MV), moderator band (MB), right atrium (RA), right ventricle (RV), tricuspid valve (TV), and spine, with corresponding textual descriptions for each structure. Data annotation was conducted by two of our co-authors with clinical backgrounds: one gynecologist with more than 5 years of experience in fetal ultrasound, and one pediatrician with 7 years of clinical and cardiological experience, including fetal cardiac imaging. After a joint induction session to agree on the labeling conventions, images were randomly sampled into two separate subsets, and each expert independently annotated one subset. We obtained ethical approval for research involving secondary data under approval no. ER-LRR-110002115202601021755 18.

The Oriented Cell Dataset (OCD) contains 120 bright-field microscopy images with oriented bounding-box annotations for a single cell class. The images and annotations are publicly distributed through the authors' dataset repository. We retained the original 24-image test set and divided the remaining 96 images into 72 training images and 24 validation images based on acquisition groups. Acquisition groups do not overlap across the training, validation, and test partitions. The resulting partitions contain 3610, 921, and 758 cell annotations, respectively.

The Ajman University Orthopantomography (AU-OPG) dataset is publicly available on Hugging Face and contains 901 panoramic dental radiographs with 7006 tooth-level oriented annotations. The images were collected at the Ajman University dental clinic, and the annotations support dental diagnosis and treatment planning. We used six diagnostic

classes: Cavitated, Crowned, Filled, Impacted, Implant, and Retained Root. The released training, validation, and test partitions contain 630, 135, and 136 radiographs, with 4964, 1010, and 1032 annotations, respectively. These predefined partitions were retained in our experiments. Radiographs from the same participant are confined to a single partition.

Table 1 summarizes the partitions.

Table 1. Dataset image partitions in training/validation/test order.

Dataset	Modality	Images	Classes
FOCUS	Fetal cardiac ultrasound	200/50/50	15
OCD	Cell microscopy	72/24/24	1
AU-OPG	Panoramic dental radiography	630/135/136	6

3.1.2. Image and Annotation Processing

Figure 1 shows representative images and ground-truth annotations from the three datasets. The following deterministic procedure prepares each image and its annotations.

1. Image conversion and bounding-box representation. Images are converted to RGB, with grayscale intensities replicated across three channels. Each annotated quadrilateral is converted to its minimum-area enclosing rectangle for training. We write an oriented bounding box as $\boldsymbol{b} = (c_x, c_y, w, h, \theta)$, where (c_x, c_y) is the rectangle center in image pixels, w and h are its width and height in pixels, and θ is the orientation of its width axis relative to the positive horizontal axis, measured in radians. We use $w \geq h$ and $\theta \in [-\pi/2, \pi/2)$. Image coordinates increase rightward and downward, so a positive angle denotes clockwise rotation. Original quadrilateral annotations were retained for evaluation.
2. Resizing and padding. For an image of width W and height H, the nominal scale is $s = \min(1024/W, 1024/H)$. Bilinear interpolation produces dimensions $W' = \lfloor sW + 0.5 \rfloor$ and $H' = \lfloor sH + 0.5 \rfloor$. Here, $\lfloor \cdot \rfloor$ denotes the floor operation, which returns the greatest integer not exceeding its argument. The resized image is placed at the upper-left corner of a 1024×1024 canvas, with zero padding on the right and bottom. The realized scale factors are $s_x = W'/W$ and $s_y = H'/H$.
3. Annotation transformation. The following implemented update expresses the training annotations in the resized image coordinate system using the realized horizontal and vertical scale factors. Right and bottom padding leave the origin unchanged, so no translation is added:

$$\begin{aligned} c'_x &= s_x c_x, \qquad c'_y = s_y c_y, \\ w' &= w\sqrt{s_x^2 \cos^2\theta + s_y^2 \sin^2\theta}, \\ h' &= h\sqrt{s_x^2 \sin^2\theta + s_y^2 \cos^2\theta}, \\ \theta' &= \operatorname{atan2}(s_x \sin\theta, s_y \cos\theta). \end{aligned} \tag{1}$$

   The transformed sides and angles are then canonicalized to the long-side convention. When integer rounding gives unequal scale factors, an affine transformation of a rotated rectangle generally produces a parallelogram, so this update is not an exact affine polygon transformation.
4. Normalization. Padded RGB values are converted to $[0, 1]$. The detector applies channel means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. These fixed RGB normalization constants follow the Torchvision preprocessing specification for ImageNet-pretrained ResNet-50 weights; they were not estimated from the medical datasets. Each channel is normalized by subtracting its corresponding mean and

dividing by its corresponding standard deviation. The same processing is used during training, validation, and testing, with no additional random augmentation.

5. Coordinate recovery. Predicted rectangles are converted to polygon vertices. Each vertex (x', y'), expressed in pixels in the resized and padded input coordinate system, is mapped to $(x'/s_x, y'/s_y)$ in the original image, where it is evaluated against the original quadrilateral annotation. Because padding is added only on the right and bottom, coordinate recovery requires no translation.

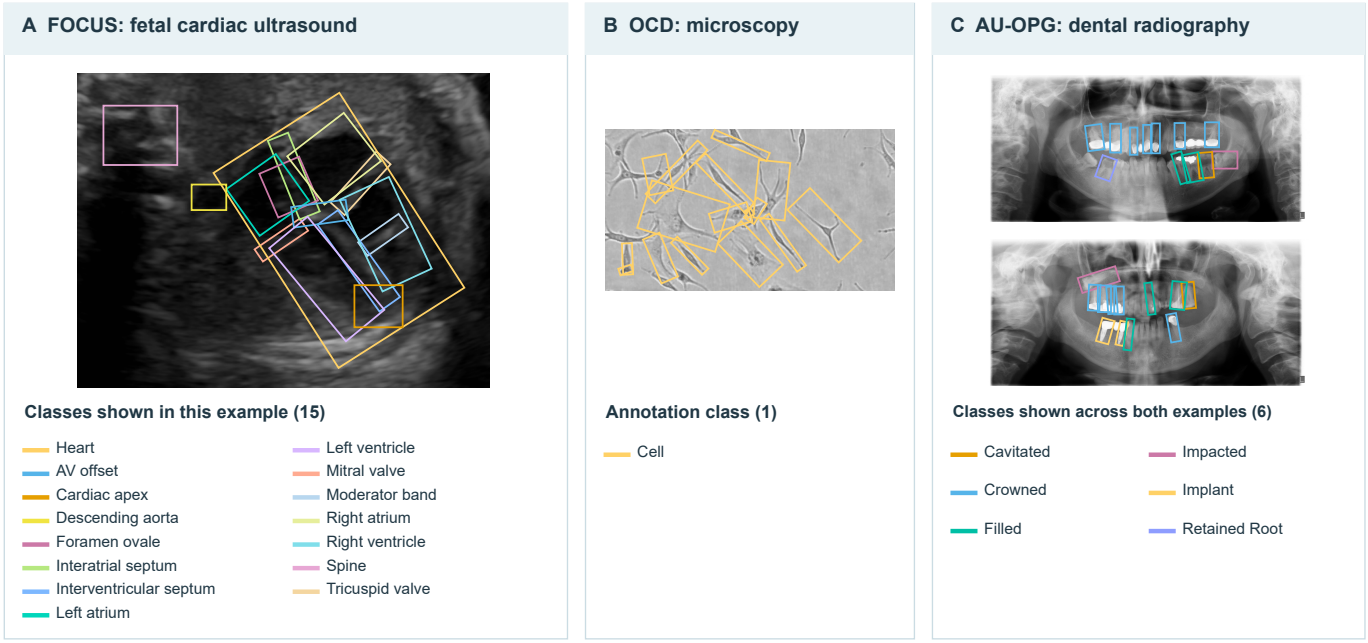


Figure 1. Sample images with ground-truth annotations from (**A**) the FOCUS dataset, (**B**) the OCD dataset, and (**C**) the AU-OPG dataset.

3.2. Overall Workflow

Figure 2 presents the GRIT-OBB method. Given a preprocessed image $\boldsymbol{I}$, a ResNet-50 backbone, and a feature pyramid network (FPN) produce multiscale features $\{\boldsymbol{F}_\ell\}$ [31,32]. A horizontal region proposal network (RPN) identifies candidate regions [33]. Each proposal $\boldsymbol{r}_i^{(0)}$ initializes a geometry learner that predicts an oriented bounding box $\widehat{\boldsymbol{b}}_i^{(1)}$. In Figure 2, these are the horizontal references after the RPN and the oriented references produced by the geometry learner, respectively.

The corrected prediction defines the sampling reference $\boldsymbol{r}_i^{(1)} = \mathrm{sg}(\widehat{\boldsymbol{b}}_i^{(1)})$, where sg denotes the stop-gradient operation. Rotated position-sensitive alignment extracts a regional feature tensor $\boldsymbol{A}_i$ in the local coordinate system defined by this reference. The RoI head maps $\boldsymbol{A}_i$ to a regional feature $\boldsymbol{z}_i$. The geometry-guided relation residual adds contextual information from other regions to obtain the context-enhanced regional feature $\boldsymbol{z}'_i$. The classification and bounding-box refinement heads use $\boldsymbol{z}'_i$ to predict the class probabilities $\boldsymbol{p}_i$ and refine the reference into the final oriented bounding box $\widehat{\boldsymbol{b}}_i^{(2)}$. The computation is summarized as follows:

$$\boldsymbol{I} \longrightarrow \{\boldsymbol{F}_\ell\} \longrightarrow \boldsymbol{r}_i^{(0)} \longrightarrow \widehat{\boldsymbol{b}}_i^{(1)} \longrightarrow \boldsymbol{A}_i \longrightarrow \boldsymbol{z}_i \longrightarrow \boldsymbol{z}'_i \longrightarrow (\boldsymbol{p}_i, \widehat{\boldsymbol{b}}_i^{(2)}). \tag{2}$$

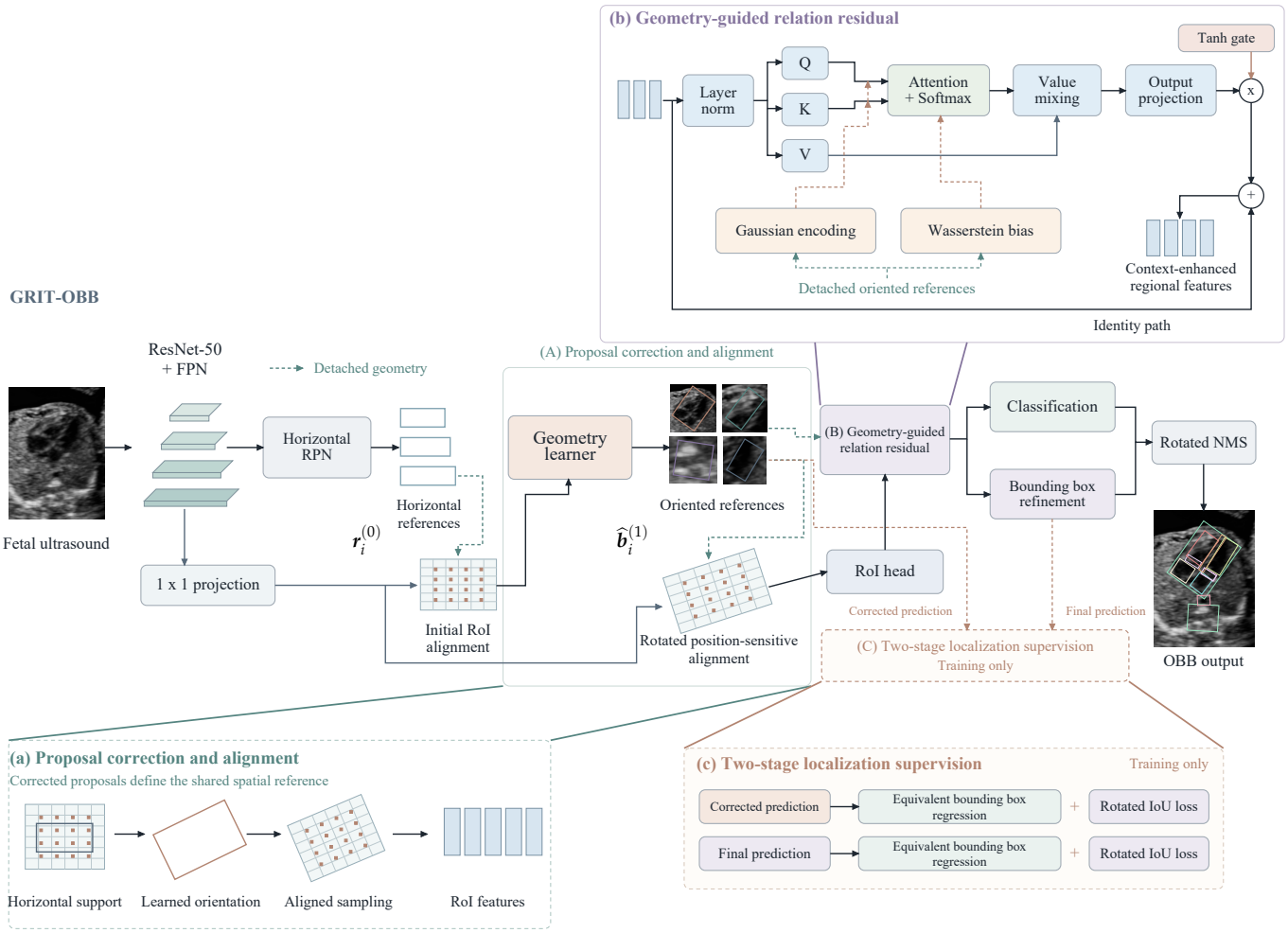


Figure 2. Overview of the GRIT-OBB method. The central diagram shows the detection workflow, with modules (**A**–**C**) expanded in panels (**a**–**c**): (**a**) proposal correction and alignment; (**b**) geometry-guided relation residual; and (**c**) two-stage localization supervision, used only during training. Colored solid guidelines connect each module to its detail panel. Teal dashed arrows indicate detached geometry, while brown dashed arrows connect the corrected and final predictions to their supervision.

The corrected reference determines the image support used for feature extraction, provides spatial information for regional interactions, and defines the coordinate frame for the final regression. The corrected proposals are detached before the second alignment to separate the two stages' geometric optimization. The geometry learner is trained by its own regression and rotated IoU losses, while the second-stage losses optimize feature extraction and final prediction without backpropagating through the corrected proposal coordinates. This design allows the corrected geometry to guide regional sampling while assigning direct localization supervision to each stage. The method is trained jointly: both stages update the shared backbone and feature pyramid, although gradients are stopped along the proposal-coordinate path. Allowing second-stage gradients to propagate through the corrected proposal coordinates could further adapt the sampling regions to the final classification and localization objectives. Such end-to-end optimization could align proposal correction more closely with the features needed for final detection, complementing the direct first-stage geometric supervision. This potential benefit has not been evaluated in the present experiments.

3.3. Geometry Learner

The supervised geometry learner converts a coarse horizontal region into an oriented reference for the subsequent feature-extraction stage. This follows the supervised region transformation principle of RoI Transformer [10]. A horizontal proposal is represented as $\boldsymbol{r} = (\boldsymbol{c}, w, h, 0)$, retaining its horizontal and vertical side lengths. Here, $\boldsymbol{c} = (c_x, c_y)^\top$ is the reference center in image pixels. Initial RoI alignment extracts a $10 \times 7 \times 7$ tensor from the proposal, and a linear layer maps its 490 flattened values to $\boldsymbol{\delta} = (\delta_x, \delta_y, \delta_w, \delta_h, \delta_\theta)$. The predicted components δ_x and δ_y are center offsets along the reference width and height axes, normalized by the corresponding side lengths; δ_w and δ_h are logarithmic side-length ratios, and δ_θ is an angular offset in radians.

The decoder converts relative updates into an oriented bounding box: the rotation matrix maps the scaled local center offsets to image coordinates, exponentiation converts logarithmic size updates to positive side lengths, and the angular update determines the new orientation. The decoded center is $\widehat{\boldsymbol{c}}$, and the resulting prediction is $\widehat{\boldsymbol{b}}$ after canonicalization. Both correction and final refinement use the same local-frame decoder:

$$\begin{aligned} \widehat{\boldsymbol{c}} &= \boldsymbol{c} + \boldsymbol{R}(\theta) \begin{bmatrix} w\delta_x \\ h\delta_y \end{bmatrix}, \\ \widehat{w} &= w \exp(\delta_w), \qquad \widehat{h} = h \exp(\delta_h), \\ \widehat{\theta} &= \mathrm{wrap}_\pi(\theta + \delta_\theta), \\ \widehat{\boldsymbol{b}} &= \mathrm{canon}(\widehat{\boldsymbol{c}}, \widehat{w}, \widehat{h}, \widehat{\theta}). \end{aligned} \tag{3}$$

Here,

$$\boldsymbol{R}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{4}$$

and wrap_π maps an angle to $[-\pi/2, \pi/2)$. The canonicalization operation exchanges the side lengths and adds $\pi/2$ when needed to maintain $w \geq h$. Side-length log updates are clipped to $[-\log(1000), \log(1000)]$, and decoded sides are lower bounded by 0.01 pixels.

The FPN supplies 256-channel maps at strides $4, 8, 16, 32$, and 64. The RPN uses anchor sizes $(16, 32)$, $(32, 64)$, $(64, 128)$, $(128, 256)$, and $(256, 512)$ at these levels, with aspect ratios 0.5, 1, and 2. The RPN retains up to 6000 proposals before horizontal non-maximum suppression (NMS) and up to 1000 afterward, using an IoU threshold of 0.7. For position-sensitive feature extraction, we use a 7×7 grid with 10 channels per bin, giving $7^2 \times 10 = 490$ output channels. A shared 1×1 convolution projects the four finer FPN levels to these position-sensitive feature maps, which are used by both regional sampling stages.

During training, horizontal envelopes of ground-truth OBBs are appended to the proposals. First-stage positives are assigned using a horizontal IoU of at least 0.5, with up to 128 positives contributing to correction supervision. Although the assignment uses horizontal envelopes, the regression targets are oriented bounding boxes. The learned correction, therefore, changes the sampling geometry before the second regional representation is constructed. Ground-truth oriented bounding boxes are also appended to the second-stage reference set during training.

3.4. Rotated Position-Sensitive Alignment

The corrected reference defines a local coordinate system for extracting image evidence. Rotated position-sensitive alignment preserves the spatial organization of this evidence

within a fixed-size regional tensor [10,34]. For a reference $\boldsymbol{r}$ with side lengths w and h, the pyramid level is

$$\ell(\boldsymbol{r}) = \mathrm{clip}_{[2,5]}\left(\left\lfloor 4 + \log_2 \frac{\sqrt{wh}}{224} \right\rfloor\right). \tag{5}$$

Here, $\lfloor \cdot \rfloor$ denotes rounding down to the nearest integer, and $\mathrm{clip}_{[2,5]}$ restricts the selected pyramid level to the range from 2 to 5. Level selection uses the reference area before context enlargement, linking feature resolution to target extent. The selected level is therefore a deterministic function of the reference, not an independent input; the dependence of the aligned feature on the level is implicit in its reference argument.

Each region is divided into $K \times K$ bins, with $K = 7$. Each bin contains S^2 sampling points, with $S = 2$. For bin row u and column v, with $u, v \in \{0, \ldots, K-1\}$, and within-bin indices $a, b \in \{0, 1\}$, define

$$\begin{aligned} \xi_{v,b} &= \frac{v + (b + 1/2)/S}{K} - \frac{1}{2}, \\ \eta_{u,a} &= \frac{u + (a + 1/2)/S}{K} - \frac{1}{2}, \\ \boldsymbol{p}_{u,v,a,b}(\boldsymbol{r}) &= \boldsymbol{c} + \boldsymbol{R}(\theta) \begin{bmatrix} \alpha_w w \xi_{v,b} \\ \alpha_h h \eta_{u,a} \end{bmatrix}. \end{aligned} \tag{6}$$

Here, $\xi_{v,b}$ and $\eta_{u,a}$ are normalized local sampling coordinates along the width and height axes, and $\boldsymbol{p}_{u,v,a,b}(\boldsymbol{r})$ is the corresponding sampling point in image coordinates. The half-step offsets place samples at the centers of equally spaced subdivisions within each bin. Scaling by the context-enlarged side lengths, rotating by the reference orientation, and adding the reference center map this regular local grid to image coordinates. For initial horizontal sampling, we set $(\alpha_w, \alpha_h) = (1, 1)$. For correctly oriented sampling, we follow the contextual RoI enlargement setting of RoI Transformer [10] and set $(\alpha_w, \alpha_h) = (1.2, 1.4)$, expanding the sampling region along the long and short axes, respectively. This enlargement incorporates the surrounding image context without changing the reference used for regression.

Let $\widetilde{\boldsymbol{F}}_\ell$ denote the projected feature map, $d_\ell = 2^\ell$ is its stride, and $k(u, v, c) = (uK + v)10 + c$ the position-sensitive channel index, where $c \in \{0, \ldots, 9\}$. The aligned tensor is

$$A_{c,u,v}(\boldsymbol{r}) = \frac{1}{S^2} \sum_{a=0}^{S-1} \sum_{b=0}^{S-1} \mathcal{B}\left(\widetilde{\boldsymbol{F}}_{\ell(\boldsymbol{r}),k(u,v,c)}, \frac{\boldsymbol{p}_{u,v,a,b}(\boldsymbol{r})}{d_{\ell(\boldsymbol{r})}} - \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}\right), \tag{7}$$

where $\mathcal{B}$ denotes bilinear interpolation with zero values outside the feature map. The element $A_{c,u,v}(\boldsymbol{r})$ is the aligned feature value for channel c and bin (u, v) of reference $\boldsymbol{r}$. The offset subtracts half a pixel from each coordinate to follow the feature-center convention. Averaging is performed separately for each bin and channel. Each bin reads its own group of 10 projected channels through $k(u, v, c)$, and each output element averages the four sampled values in that bin. This forms a fixed-size regional representation while retaining the association between feature channels and spatial bins. For bilinear interpolation, let the fractional horizontal and vertical offsets from the lower integer coordinates be t_x and t_y. The four neighboring values receive weights $(1 - t_x)(1 - t_y)$, $t_x(1 - t_y)$, $(1 - t_x)t_y$, and $t_x t_y$, respectively. For example, values 2, 4, 6, and 8 at the upper-left, upper-right, lower-left, and lower-right neighbors yield 4.5 when $t_x = 0.25$ and $t_y = 0.5$. An out-of-bounds neighbor contributes zero; the remaining weights are not renormalized.

The RoI head flattens the resulting $10 \times 7 \times 7$ tensor and applies a fully connected layer, layer normalization, and ReLU to obtain $\boldsymbol{z}_i \in \mathbb{R}^{1024}$. This regional feature represents

the appearance in the corrected frame and supplies input to geometric interaction and the final prediction.

3.5. Geometry-Guided Relation Residual

Regional interaction allows an aligned target representation to incorporate information from other proposals. The geometry-guided relation residual, labeled Geometry-guided relation residual in the upper workflow of Figure 2, uses Gaussian positional encoding and Wasserstein geometric relationships [14]. Each detached reference defines a Gaussian distribution:

$$\boldsymbol{\mu}_i = \boldsymbol{c}_i, \qquad \Sigma_i = \boldsymbol{R}(\theta_i)\,\mathrm{diag}(w_i^2/4, h_i^2/4)\boldsymbol{R}(\theta_i)^{\mathsf{T}}. \tag{8}$$

Here, i indexes a reference region, $\boldsymbol{\mu}_i \in \mathbb{R}^2$ is its mean, and $\Sigma_i \in \mathbb{R}^{2\times 2}$ is its covariance matrix. The mean describes location, while the covariance encodes extent and orientation. For two regions, the Gaussian Wasserstein distance and normalized attention bias are

$$\begin{aligned} C_{ij} &= \left[\mathrm{tr}(\Sigma_i) + \mathrm{tr}(\Sigma_j) - 2\sqrt{\mathrm{tr}(\Sigma_i\Sigma_j) + 2\sqrt{\det(\Sigma_i)\det(\Sigma_j)}}\right]_+, \\ D_{ij} &= \sqrt{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2^2 + C_{ij}}, \\ B_{ij} &= -\log\left(1 + \frac{D_{ij}}{(w_i h_i w_j h_j)^{1/4}}\right). \end{aligned} \tag{9}$$

Here, i and j index regions in the same image; C_{ij} is the covariance contribution to the squared Gaussian Wasserstein distance, D_{ij} is the distance, and B_{ij} is the normalized attention bias. The operators tr and det denote the matrix trace and determinant, and $\|\cdot\|_2$ denotes the Euclidean norm. The denominator normalizes distance by the spatial scale of the pair, and $B_{ii} = 0$. We use $[x]_+ = \max(x, 0)$ to suppress negative numerical round-off.

Gaussian positional encoding uses the expected sinusoidal features of the coordinate marginals. For axis $d \in \{x, y\}$ and $\omega_k = 10{,}000^{-k/32}$, $k = 0, \ldots, 31$,

$$\begin{aligned} \phi^{\sin}_{i,d,k} &= \exp(-\omega_k^2\Sigma_{i,dd}/2)\sin(\omega_k\mu_{i,d}), \\ \phi^{\cos}_{i,d,k} &= \exp(-\omega_k^2\Sigma_{i,dd}/2)\cos(\omega_k\mu_{i,d}). \end{aligned} \tag{10}$$

Here, $\mu_{i,d}$ and $\Sigma_{i,dd}$ denote the mean and variance of proposal i along axis d, respectively, and k indexes the sinusoidal frequencies. Concatenation over both axes gives a 128-dimensional encoding. A learned affine projection maps each layer-normalized regional feature from 1024 to 384 dimensions, which are split into separate 128-dimensional query, key, and value vectors. Each vector is then divided into four 32-dimensional attention heads. Queries and keys determine the attention weights between regions, while values supply the regional information aggregated using those weights. Positional encoding enters the queries and keys, and the distance bias conditions their interaction:

$$\begin{aligned} a_{ij}^{(m)} &= \mathrm{softmax}_j\left[\frac{(\boldsymbol{q}_i^{(m)} + \boldsymbol{\phi}_i^{(m)})^{\mathsf{T}}(\boldsymbol{k}_j^{(m)} + \boldsymbol{\phi}_j^{(m)})}{\sqrt{32}} + \mathrm{softplus}(\tau_m)B_{ij}\right], \\ \boldsymbol{z}_i' &= \boldsymbol{z}_i + \tanh(g)\left[\boldsymbol{W}_o\,\mathrm{Concat}_{m=1}^{4}\left(\sum_j a_{ij}^{(m)}\boldsymbol{v}_j^{(m)}\right) + \boldsymbol{b}_o\right]. \end{aligned} \tag{11}$$

Here, $m \in \{1, 2, 3, 4\}$ indexes attention heads, and $\boldsymbol{q}_i^{(m)}, \boldsymbol{k}_i^{(m)}, \boldsymbol{v}_i^{(m)} \in \mathbb{R}^{32}$ are the query, key, and value vectors for region i. The vector $\boldsymbol{\phi}_i^{(m)}$ is the corresponding 32-dimensional part of the concatenated positional encoding, and $a_{ij}^{(m)}$ is the attention weight from region

i to region j. Both softmax_j and the sum over j operate over the reference regions in the same image. Each head has its own learnable scalar τ_m, optimized jointly with the feature projections. Its positive scale $\text{softplus}(\tau_m)$ weights the geometric bias B_{ij} relative to the query-key similarity, where $\text{softplus}(x) = \log(1 + e^x)$. The learnable scalar g controls the residual gate through $\tanh(g)$. The operator Concat concatenates the four head outputs. The four 32-dimensional head outputs are concatenated into a 128-dimensional vector. The output affine layer uses the learned weight matrix $\boldsymbol{W}_o \in \mathbb{R}^{1024\times128}$ and bias vector $\boldsymbol{b}_o \in \mathbb{R}^{1024}$. It restores the feature dimension to 1024 so that the gated result can be added to the original regional feature. The effective bias scale is initialized to 1, and the residual gate is initialized to $\tanh(g) = 0.1$. Reference geometry and its positional encoding remain detached; feature projections, bias scales, and the gate are learned. The residual structure retains the aligned regional feature while adding geometry-conditioned contextual information.

3.6. Two-Stage Localization Supervision

3.6.1. Class and Bounding-Box Prediction

The context-enhanced regional feature $\boldsymbol{z}'_i$ feeds a classifier with $C + 1$ outputs, where C is the number of foreground classes and the additional output represents background, and a class-agnostic regressor with five outputs. The regressor uses Equation 3 relative to the corrected reference. Second-stage positives are assigned by a rotated IoU of at least 0.5. At most 512 regions are sampled, with a maximum positive fraction of 0.25. Classification uses all sampled regions, while bounding-box regression uses the positive subset.

3.6.2. Equivalent Bounding-Box Regression

An oriented rectangle admits equivalent parameterizations: (w, h, θ) and $(h, w, \theta + \pi/2)$. The regression objective accounts for this equivalence when comparing predicted updates with target codes. For reference $\boldsymbol{r} = (\boldsymbol{c}_r, w_r, h_r, \theta_r)$ and target $\boldsymbol{b} = (\boldsymbol{c}_b, w_b, h_b, \theta_b)$, define

$$\boldsymbol{E}_{\boldsymbol{r}}(\boldsymbol{b}) = \begin{bmatrix} \text{diag}(w_r^{-1}, h_r^{-1})\boldsymbol{R}(\theta_r)^{\mathsf{T}}(\boldsymbol{c}_b - \boldsymbol{c}_r) \\ \log(w_b/w_r) \\ \log(h_b/h_r) \\ \text{wrap}_\pi(\theta_b - \theta_r) \end{bmatrix}. \tag{12}$$

The vector $\boldsymbol{E}_{\boldsymbol{r}}(\boldsymbol{b}) \in \mathbb{R}^5$ encodes the target bounding box $\boldsymbol{b}$ relative to reference $\boldsymbol{r}$, in center, side-length, and angle coordinates. The center difference is rotated into the reference frame and normalized by its side lengths; logarithmic size ratios and the wrapped angular difference complete the relative target. These five quantities match the decoder's predicted updates, allowing regression to compare predictions and targets in the same coordinate system. Let $\boldsymbol{b}^{[0]}$ and $\boldsymbol{b}^{[1]}$ denote the two equivalent target representations. The per-positive regression loss is

$$\mathcal{L}_{\text{code}}(\boldsymbol{\delta}; \boldsymbol{r}, \boldsymbol{b}) = \min_{j\in\{0,1\}} \sum_{k=1}^{5} \text{SmoothL1}_{1/9}\Big(\mathcal{W}_k[\boldsymbol{\delta} - \boldsymbol{E}_{\boldsymbol{r}}(\boldsymbol{b}^{[j]})]\Big). \tag{13}$$

Here, $\boldsymbol{\delta} \in \mathbb{R}^5$ is the predicted regression vector, $j \in \{0, 1\}$ indexes the equivalent target representations, and $k \in \{1, \ldots, 5\}$ indexes regression coordinates. The loss $\text{SmoothL1}_{1/9}$ uses transition parameter $\beta = 1/9$. The operator $\mathcal{W}_k$ returns the kth residual component, leaving center and side-length differences unchanged and wrapping the angular difference with period π. The minimum is taken after summing the five coordinate penalties, allowing either equivalent target representation to determine the regression error. This avoids penalizing a geometrically identical rectangle solely because its width and height are

exchanged with a corresponding quarter-turn in angle. Smooth L1 is quadratic for residual magnitudes below β and linear beyond that threshold, reducing the influence of large coordinate errors relative to a purely quadratic penalty. The per-positive loss is shared by both stages, so its definition has no stage superscript. For stage s, the term $\overline{\mathcal{L}}_{\text{code}}^{(s)}$ is the mean of this same loss over the sampled positive regions at that stage, using their predicted updates, references, and assigned targets. Stage 1 uses horizontal references, whereas stage 2 uses detached corrected references; an empty positive set contributes zero.

3.6.3. Direct Overlap Supervision

Coordinate regression is supplemented by direct overlap supervision at both geometric stages. Coordinate errors and spatial overlap measure different aspects of localization. The rotated IoU loss directly penalizes disagreement between the predicted and assigned target regions, complementing the parameter-space regression loss. Applying it to both stages ensures the corrected sampling reference and the final detection are supervised. For the positive set $\mathcal{P}_s$ of stage $s \in \{1, 2\}$,

$$\begin{aligned} \mathcal{L}_{\text{IoU}}^{(s)} &= \frac{1}{|\mathcal{P}_s|} \sum_{i \in \mathcal{P}_s} \left[1 - \text{rIoU}(\widehat{\boldsymbol{b}}_i^{(s)}, \boldsymbol{b}_i^*) \right], \\ \text{rIoU}(\boldsymbol{a}, \boldsymbol{b}) &= \frac{|P(\boldsymbol{a}) \cap P(\boldsymbol{b})|}{|P(\boldsymbol{a}) \cup P(\boldsymbol{b})|}, \end{aligned} \tag{14}$$

Here, $s = 1$ denotes proposal correction and $s = 2$ denotes final bounding-box refinement. The symbol $\boldsymbol{b}_i^*$ denotes the ground-truth bounding box assigned to prediction $\widehat{\boldsymbol{b}}_i^{(s)}$, and $|\mathcal{P}_s|$ is the number of positive regions at stage s. For any bounding boxes $\boldsymbol{a}$ and $\boldsymbol{b}$, $P(\boldsymbol{a})$ and $P(\boldsymbol{b})$ denote their rectangle polygons; $|\cdot|$ in the intersection-over-union ratio denotes polygon area. An empty positive set contributes zero. Intersection vertices are obtained from contained corners and edge intersections, ordered around their centroid, and used to compute polygon area. The resulting overlap objective is piecewise differentiable through the selected coordinates.

The training objective combines proposal generation, class prediction, and localization: the RPN terms train foreground scoring and horizontal proposal regression, the classification term trains the final class predictions, and the two pairs of regression and overlap terms supervise proposal correction and final refinement. These terms are optimized jointly, with unit coefficients except for the two overlap terms shown below. The complete training objective is

$$\begin{aligned} \mathcal{L} = &\mathcal{L}_{\text{RPN,obj}} + \mathcal{L}_{\text{RPN,reg}} + \overline{\mathcal{L}}_{\text{code}}^{(1)} + \mathcal{L}_{\text{cls}} + \overline{\mathcal{L}}_{\text{code}}^{(2)} \\ &+ \lambda_1 \mathcal{L}_{\text{IoU}}^{(1)} + \lambda_2 \mathcal{L}_{\text{IoU}}^{(2)}. \end{aligned} \tag{15}$$

Here, $\mathcal{L}_{\text{RPN,obj}}$ and $\mathcal{L}_{\text{RPN,reg}}$ are the RPN objectness classification and bounding-box regression losses, respectively. Superscripts (1) and (2) denote proposal correction and final bounding-box refinement. Bars denote averages over the corresponding positive regions, and $\mathcal{L}_{\text{cls}}$ is softmax cross-entropy over sampled second-stage regions. RPN objectness is the mean binary cross-entropy with logits over sampled positive and negative anchors. Up to 256 anchors are sampled per image, with at most half positive. RPN regression sums Smooth L1 penalties with $\beta = 1/9$ over the four encoded coordinates of positive anchors and divides by the total number of sampled positive and negative anchors. Its targets are the horizontal envelopes of the assigned ground-truth rectangles. Final classification averages cross-entropy over all sampled second-stage regions across the C foreground classes and one background class; background regions do not contribute to the localization losses. The first overlap term trains the corrected sampling reference; the second trains the final detection. The coefficients λ_1 and λ_2 control direct overlap supervision for proposal

correction and final refinement, respectively. We used $\lambda_1 = \lambda_2 = 2$ as the default setting, giving equal weight to the two stages.

The stage boundary separates the geometric gradient paths: the geometry learner is updated by its own regression and overlap losses, and the final heads are updated by second-stage objectives. Both stages also train the shared visual features. Horizontal IoU is used for the first-stage assignment; rotated polygon IoU is used for the second-stage assignment and the direct overlap losses. Gaussian bounding-box representations are used only in the regional interaction module.

4. Experimental Evaluation

This section evaluates whether geometry-guided feature extraction and two-stage localization supervision improve fetal cardiac structure detection. We first describe the experimental settings and compare GRIT-OBB with six oriented detectors on the FOCUS dataset, examining overall accuracy, class-level performance, and qualitative examples. We then evaluate the method on OCD and AU-OPG datasets, assess robustness to geometric and visibility perturbations, and use component ablations to examine the contributions of the proposed design.

4.1. Experimental Settings

4.1.1. Compared Models and Training Procedure

We compared GRIT-OBB with six oriented detectors. Rotated RetinaNet, R3Det, and S^2A-Net represent dense prediction, feature refinement, and feature alignment, respectively [11,12,18]. Oriented-DETR and O2-RT-DETR provide transformer-based comparisons [21,22], while GauCho-FCOS uses an anchor-free detector with a Gaussian bounding box representation [19,24].

Each model was trained independently five times on each dataset using the same fixed training, validation, and test partitions. All models used the same input resolution, batch size, and maximum training budget listed in Table 2, while retaining their model-specific architectures and optimizers. Checkpoints were selected by validation, rotated AP50, and operating thresholds were determined on the validation set before test evaluation.

GRIT-OBB uses an ImageNet-pretrained ResNet-50 backbone with a 256-channel feature pyramid. Training used AdamW [35] and cosine learning-rate decay over the 100-epoch budget. Table 2 summarizes the principal training parameters.

Table 2. Principal experimental settings. Input size, batch size, maximum epochs, and early-stopping settings are shared across models; the optimizer and learning-rate settings shown here apply to GRIT-OBB. Early stopping was monitored starting at epoch 10 using the validation rotated AP50, with a minimum improvement of 0.001.

Parameter	Value	Parameter	Value
Input size	1024×1024	Batch size	4
Maximum epochs	100	Early-stopping patience	8
Optimizer	AdamW	Learning-rate schedule	Cosine decay
Initial learning rate	10^{-4}	Weight decay	10^{-4}

We additionally evaluated GRIT-OBB using image-level five-fold cross-validation over all 300 FOCUS images. Each fold contained 60 test images, with the remaining 240 divided into 192 training and 48 validation images. Each image served as a test sample once within the five-fold partition. Checkpoints were selected using the AP50 validation set, and operating thresholds were determined on the corresponding validation partition

prior to test evaluation. We report the mean and sample standard deviation across the five folds.

4.1.2. Evaluation Measures

Detection was evaluated using rotated polygon IoU. We report class-averaged AP50, AP75, and mAP50:95, together with AR@100. The latter two measures average over IoU thresholds from 0.50 to 0.95 in 0.05 increments; AR@100 uses the top 100 detections per image and pools ground-truth recall.

Micro-F1 and mean matched rotated IoU were evaluated on the test set at a fixed score threshold selected on the validation set for each training run, using a rotated IoU threshold of 0.5.

A pooled optimal LRP-style error, $\mathrm{oLRP}_{\mathrm{pool}}$, summarizes localization, false positives, and false negatives [36]. At each score threshold, the unnormalized overlap errors $1 - \mathrm{IoU}_{\mathrm{rot}}$ of true-positive matches are added to the false-positive and false-negative counts and divided by the total number of true positives, false positives, and false negatives. Counts are pooled across images and classes, and the minimum over thresholds is reported. Lower $\mathrm{oLRP}_{\mathrm{pool}}$ is better; higher values are better for the other detection measures.

4.1.3. Robustness Evaluation and Statistical Analysis

Robustness was evaluated under 12 test conditions: rotations of 5°, 10°, 20°, and 30°; scale factors of 0.85, 0.70, 0.55, and 0.40; and target-aware occlusion fractions of 0.10, 0.20, 0.30, and 0.40. Image content and annotations were transformed together for rotation and scaling. All models used fixed checkpoints and the same perturbed inputs. Rotation was evaluated on annotations fully contained within the common image region that remains within the canvas under all tested rotations, in either direction. The clean reference used the same target subset, termed the rotation core. Scaling and occlusion used the full clean reference. Robustness preprocessing was shared across models, with matching clean evaluations for normalization.

We summarize absolute robustness by r-mAP, the mean mAP50:95 across all 12 conditions, and by the lowest mAP50:95 among the three most severe conditions. Relative retention is measured by RR-AUC, the mean of the trapezoidal areas under the clean-normalized accuracy curves for the three perturbation families. For integration, perturbation severity ranges from 0 (clean) to 1 (the strongest tested condition). Rotation angles are normalized by the maximum tested angle of 30°, whereas scaling and occlusion use equally spaced severity levels. Each summary was computed separately for each dataset and training run.

Results are summarized by the mean and sample standard deviation of five independent training runs, with 95% Student's t confidence intervals describing training variability. Paired differences in r-mAP and RR-AUC were assessed using 10,000 hierarchical-cluster/image bootstrap resamples, with shared resampling weights across the compared models and a Holm correction over six baselines per dataset and endpoint.

4.2. Fetal Cardiac Structure Localization

4.2.1. Overall Detection Performance

The main comparison tested whether GRIT-OBB improves both target retrieval and localization. Figure 3 shows the eight metrics and the distribution across five training runs. GRIT-OBB had the best mean on all eight metrics. Table 3 separates detection and recall from F1, overlap, sensitivity, and error measures, with explicit uncertainty and first/second-place markings.

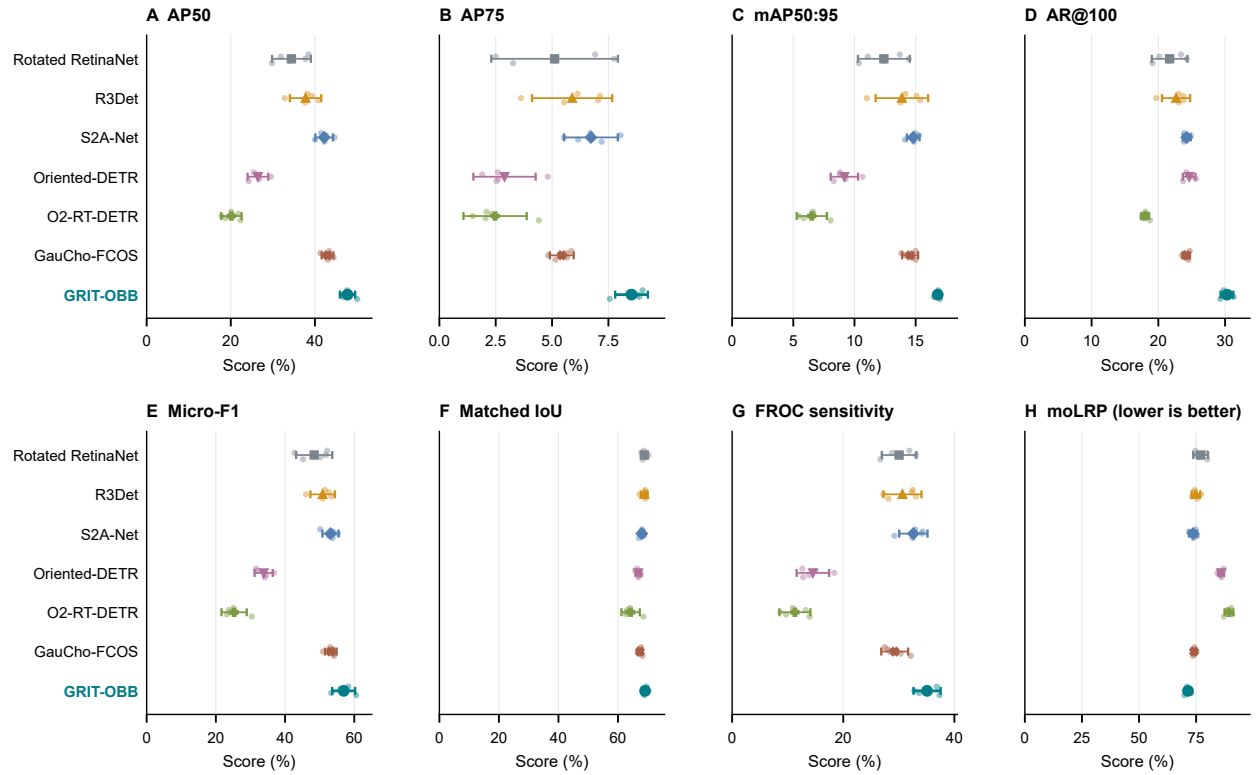


Figure 3. Performance across eight metrics on the FOCUS dataset. Plots summarize the results across the training seeds for each detector; larger symbols and horizontal intervals show five-seed means and 95% t intervals ($\bar{x} \pm 2.776s/\sqrt{5}$), respectively. All scores are percentages. Higher is better, except for moLRP, which pools optimal error estimates from localization errors, false positives, and false negatives. FROC sensitivity averages sensitivity at 0.25, 0.5, 1, and 2 false positives per image. Results are from the clean test set.

GRIT-OBB reached $47.74 \pm 1.46\%$ AP50, with a 95% interval of [45.93, 49.56]%. The strongest AP50 baseline, GauCho-FCOS, achieved 43.00%. At IoU 0.75, GRIT-OBB achieved $8.53 \pm 0.59\%$, compared with 6.73% for S^2A-Net. Its mAP50:95 reached $16.81 \pm 0.19\%$, compared with 14.80% for S^2A-Net. The improvements at both overlap thresholds show that the benefit extends from detection at IoU 0.5 to stricter localization.

Recall, F1, and FROC sensitivity results support the same pattern. AR@100 rose from the strongest baseline value of 24.62% to 30.28%, and micro-F1 rose from 53.23% to 56.94%. FROC mean sensitivity reached 35.09%, compared with 32.62% for S^2A-Net, while $oLRP_{pool}$ decreased from 73.69% to 71.43%. Mean matched IoU was 69.11% for GRIT-OBB and 68.89% for Rotated RetinaNet. Together, the measures indicate that GRIT-OBB retrieves more targets while maintaining the quality of matched detections' overlap. The low absolute AP75 across methods indicates that precise placement remains a continuing challenge on the FOCUS dataset.

Table 3. Summary results on the FOCUS dataset. The first line reports the mean ± sample SD, and the second reports the 95% training-seed t confidence interval ($n = 5$). Values in red and blue indicate the best and second-best unrounded means for each metric, respectively; rankings do not indicate statistical significance.

Model	AP50 ↑	AP75 ↑	mAP50:95 ↑	AR@100 ↑
Rotated RetinaNet	34.41 ± 3.72 [29.79, 39.02]	5.11 ± 2.26 [2.30, 7.92]	12.40 ± 1.71 [10.28, 14.52]	21.69 ± 2.15 [19.02, 24.35]
R3Det	37.77 ± 3.00 [34.04, 41.50]	5.89 ± 1.43 [4.11, 7.67]	13.87 ± 1.73 [11.73, 16.02]	22.65 ± 1.69 [20.55, 24.76]
S^2A-Net	42.19 ± 1.70 [40.09, 44.30]	6.73 ± 0.96 [5.53, 7.92]	14.80 ± 0.42 [14.28, 15.32]	24.25 ± 0.47 [23.67, 24.84]
Oriented-DETR	26.44 ± 1.98 [23.99, 28.90]	2.89 ± 1.11 [1.51, 4.27]	9.17 ± 0.90 [8.06, 10.29]	24.62 ± 0.76 [23.67, 25.56]
O2-RT-DETR	20.15 ± 1.95 [17.73, 22.56]	2.47 ± 1.13 [1.07, 3.87]	6.51 ± 0.99 [5.28, 7.75]	18.02 ± 0.47 [17.43, 18.61]
GauCho-FCOS	43.00 ± 1.13 [41.60, 44.40]	5.43 ± 0.43 [4.90, 5.96]	14.55 ± 0.53 [13.89, 15.20]	24.17 ± 0.46 [23.60, 24.73]
GRIT-OBB	47.74 ± 1.46 [45.93, 49.56]	8.53 ± 0.59 [7.80, 9.25]	16.81 ± 0.19 [16.57, 17.04]	30.28 ± 0.77 [29.32, 31.24]
Model	**Micro-F1 ↑**	**Matched IoU ↑**	**FROC Mean ↑**	**$oLRP_{pool}$ ↓**
Rotated RetinaNet	48.41 ± 4.21 [43.19, 53.64]	68.89 ± 0.91 [67.75, 70.02]	30.05 ± 2.51 [26.94, 33.17]	76.95 ± 2.59 [73.74, 80.17]
R3Det	50.86 ± 2.86 [47.31, 54.41]	68.84 ± 0.92 [67.70, 69.98]	30.65 ± 2.77 [27.22, 34.09]	74.89 ± 1.52 [73.01, 76.78]
S^2A-Net	53.14 ± 1.89 [50.79, 55.49]	68.00 ± 0.61 [67.25, 68.76]	32.62 ± 2.05 [30.08, 35.16]	73.69 ± 1.46 [71.87, 75.50]
Oriented-DETR	33.82 ± 2.13 [31.18, 36.46]	66.82 ± 0.77 [65.87, 67.77]	14.54 ± 2.35 [11.62, 17.46]	85.94 ± 1.08 [84.60, 87.27]
O2-RT-DETR	25.28 ± 2.94 [21.63, 28.92]	64.24 ± 2.48 [61.16, 67.32]	11.30 ± 2.24 [8.51, 14.08]	89.38 ± 1.55 [87.46, 91.30]
GauCho-FCOS	53.23 ± 1.33 [51.57, 54.88]	67.37 ± 0.62 [66.61, 68.14]	29.26 ± 1.95 [26.84, 31.68]	74.11 ± 0.66 [73.29, 74.93]
GRIT-OBB	56.94 ± 2.67 [53.62, 60.25]	69.11 ± 0.76 [68.17, 70.05]	35.09 ± 1.97 [32.64, 37.53]	71.43 ± 1.10 [70.07, 72.79]

4.2.2. Class-Level Localization

Figure 4 shows substantial differences in localization performance across the 15 anatomical classes. GRIT-OBB achieved the highest mean AP50 in nine classes, tied for the heart, and fell below the strongest baseline in five. The heatmap shows the absolute accuracy of each method, while the difference plot identifies the classes in which GRIT-OBB outperformed the strongest baseline.

GRIT-OBB performed best on the whole heart and cardiac chambers. It achieved 100% AP50 for the heart, matching Rotated RetinaNet, R3Det, and S^2A-Net. The right atrium reached approximately 88.2%, while the left atrium, left ventricle, and right ventricle each exceeded 70%. These results indicate stronger localization of the overall heart and chamber regions than of several finer anatomical targets. However, whole-heart detection does not distinguish GRIT-OBB from the other leading methods; its advantages are more apparent within the heart.

The largest improvements occurred for the interventricular septum, mitral valve, and right ventricle. Their AP50 values reached 55.91%, 37.03%, and 70.65%, compared with 38.11%, 23.48%, and 59.87% for the respective strongest baselines. The gains, therefore, span septal, valvular, and chamber structures. While the mitral valve showed a substantial improvement over the strongest baseline, its absolute AP50 remained well below that of the

chambers. Similarly, the interatrial septum and foramen ovale outperformed their strongest baselines, but their AP50 values remained at approximately 13.7% and 18.7%, respectively.

FOCUS | Anatomical class performance

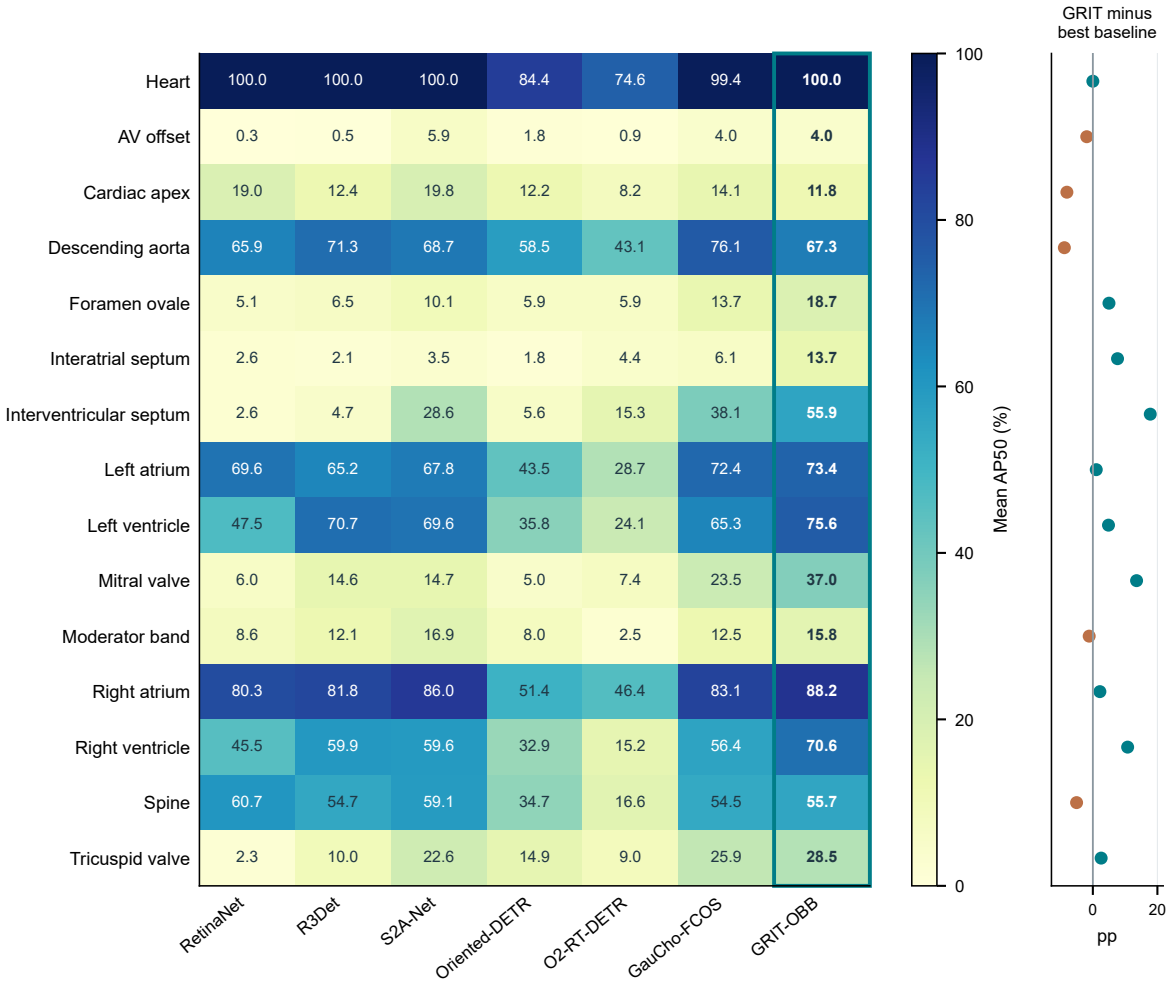


Figure 4. Class-specific localization performance on the FOCUS dataset. The heatmap shows five-seed mean AP50 for all 15 annotation classes and seven models; the GRIT-OBB column is outlined. Darker blue cells indicate higher AP50, whereas lighter yellow cells indicate lower AP50, on a common scale from 0 to 100%. The right panel shows the difference between GRIT-OBB and the highest baseline mean for each class, calculated by subtracting the baseline score from the GRIT-OBB score, with both scores expressed as percentages. The strongest baseline may differ across classes, and negative values indicate lower AP50 for GRIT-OBB.

The weakest results were observed for atrioventricular valve offset and the cardiac apex, with AP50 values of 4.01% and 11.76%, respectively. The moderator band also remained difficult, at approximately 15.8%. GRIT-OBB trailed the strongest baseline for these three classes, as well as for the descending aorta and spine. However, the latter two retained substantially higher absolute AP50 values, approximately 67.3% and 55.7%, illustrating that a relative disadvantage does not necessarily imply poor localization. Overall, the class results show strong chamber localization and substantial gains for selected internal structures, while fine anatomical landmarks and several small intracardiac targets remain the main weaknesses.

4.2.3. Qualitative Comparison

Figure 5 presents two images from the FOCUS dataset: the first row illustrates stronger GRIT-OBB localization, while the second shows remaining detection errors. Each row compares the ground truth, six baselines, and GRIT-OBB using the same image crop and class colors. The comparison examines whether predictions recover the annotated structures and match their positions, extents, and orientations.

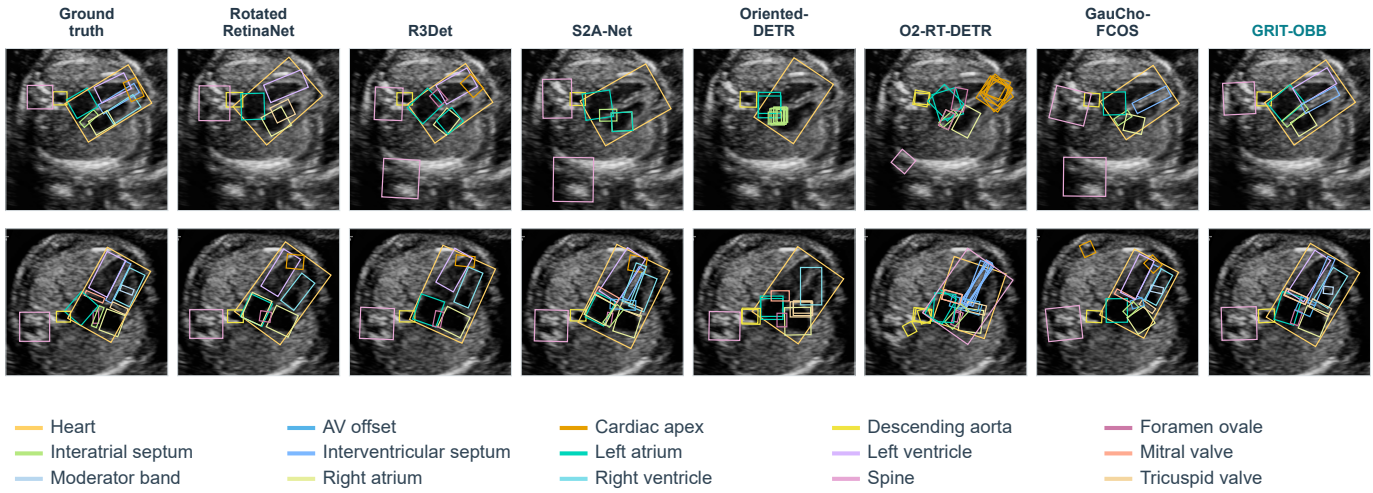


Figure 5. Qualitative comparison on the FOCUS dataset. The first row illustrates stronger overall localization, while the second highlights remaining errors in detection and localization for small anatomical structures. Columns show the ground-truth annotations, six baseline detectors, and GRIT-OBB, with consistent class colors and the same image crop within each row.

In the first row, GRIT-OBB predicted oriented bounding boxes that closely matched the ground-truth position, extent, and orientation of the whole heart. The left atrium, right atrium, and left ventricle were also localized, together with the interventricular septum, descending aorta, and spine. Several baseline detectors produced displaced or additional bounding boxes, whereas GRIT-OBB showed closer agreement with the annotations across these structures.

In the second row, GRIT-OBB successfully localized the whole heart and all four cardiac chambers, as well as the mitral valve, tricuspid valve, interventricular septum, moderator band, descending aorta, and spine. However, the interatrial septum was missed, and the oriented bounding box for the foramen ovale did not adequately overlap its annotation. Additional predictions were also present. This example illustrates localized errors in smaller anatomical structures despite accurate detection of most annotated targets.

Together, these examples show that accurate localization of the heart and cardiac chambers can coexist with missed or imprecisely localized smaller structures. They complement the class-level results by illustrating both the anatomical coverage achieved by GRIT-OBB and the remaining need to improve detection and bounding-box localization of small cardiac structures.

4.2.4. Five-Fold Cross-Validation

Table 4 reports the five-fold results on FOCUS. Mean AP50 was 41.35%, with a sample standard deviation of 1.77%; AP75 and mAP50:95 were 7.01% and 14.66%, respectively. AP50 ranged from 38.72% to 42.73% across the five test folds. These results describe performance across alternative image-level partitions of the dataset.

Table 4. Five-fold cross-validation of GRIT-OBB on FOCUS. All metrics are expressed as percentages. The final row reports the mean and sample SD across folds.

Fold	AP50	AP75	mAP50:95	AR@100
1	42.73	8.89	15.64	30.63
2	38.72	7.23	14.24	28.85
3	42.45	7.19	14.65	30.02
4	42.55	6.39	15.38	31.10
5	40.31	5.33	13.37	29.59
Mean ± SD	41.35 ± 1.77	7.01 ± 1.31	14.66 ± 0.91	30.04 ± 0.88
Fold	**Micro-F1**	**Matched IoU**	**FROC Mean**	**moLRP**
1	56.56	69.62	32.20	71.34
2	51.62	70.18	30.39	74.82
3	49.34	68.87	34.02	74.17
4	51.77	68.88	31.91	74.00
5	41.89	67.70	28.54	76.65
Mean ± SD	50.24 ± 5.35	69.05 ± 0.93	31.41 ± 2.06	74.20 ± 1.91

The cross-validation means were lower than the original fixed-partition means of 47.74% AP50, 8.53% AP75, and 16.81% mAP50:95. The two evaluations use different training and test partitions. The cross-validation results therefore provide an additional estimate of performance across image-level splits, while the fixed-partition experiment provides the comparison with the six baseline detectors.

4.2.5. Localization Error Analysis

We analyzed the full model's predictions from five training runs on the FOCUS test set, using each run's validation-selected confidence threshold. Predictions were matched to the ground truth within each class in descending order of confidence, with one-to-one matching at rotated IoU thresholds of 0.25 and 0.50. Center displacement was normalized by the ground-truth diagonal. Angular error was the smaller orientation difference modulo 180 degrees; targets with a long-to-short-side ratio below 1.1 were excluded from angle summaries because their orientation is ambiguous. Dimension errors were the absolute natural logarithms of predicted-to-ground-truth long- and short-side ratios. Small, medium, and large groups were defined using the lower and upper tertiles of image-normalized target area in the training annotations, corresponding to area fractions of 0.00514 and 0.01394.

As shown in Table 5, small targets showed the largest loss of recall when the matching threshold increased from 0.25 to 0.50, falling from 41.02% to 28.52%, compared with 86.57% to 83.58% for large targets. At IoU 0.50, small targets also had greater normalized center displacement (0.069 versus 0.045) and short-side error (0.176 versus 0.099) than large targets. Their short-side error exceeded their long-side error (0.116), indicating that estimating dimensions along the shorter axis is a particular difficulty. Angular errors among matched small and large targets were closer, at 5.80 and 5.50 degrees, respectively. This analysis identifies small-target retrieval, relative center displacement, and short-side estimation as important limitations of precise localization.

Figure 6 illustrates six small-structure failure cases from the test predictions. The tricuspid and mitral valves were missed at the validation-selected confidence threshold of 0.4644. Predictions for the foramen ovale, descending aorta, atrioventricular valve offset, and moderator band showed insufficient overlap with ground truth, with IoU values of approximately 0.39-0.45. Small targets were defined by the lower training-set tertile of normalized annotation area (0.00514). These examples illustrate missed detections and errors in position or extent; they were selected to show failure modes rather than estimate their frequency.

Table 5. Localization errors by target size on FOCUS. Values are mean and sample SD across five runs: recall is calculated per run, while geometric errors summarize the per-run median among matched targets. Lower geometric errors are better. N denotes the number of ground-truth targets. Center error is normalized by the target diagonal; long- and short-side errors are absolute log ratios. Unmatched targets contribute to recall but not to geometric-error summaries.

Matching IoU: 0.25						
Size	**N**	**Recall (%)**	**Center Error**	**Angle (°)**	**Long Side**	**Short Side**
All	455	60.44 ± 6.02	0.061 ± 0.004	6.18 ± 0.26	0.114 ± 0.007	0.129 ± 0.010
Small	176	41.02 ± 10.50	0.079 ± 0.015	6.26 ± 0.62	0.135 ± 0.017	0.217 ± 0.029
Medium	145	59.86 ± 4.40	0.066 ± 0.005	7.39 ± 2.16	0.125 ± 0.013	0.144 ± 0.017
Large	134	86.57 ± 2.53	0.047 ± 0.002	5.71 ± 0.23	0.097 ± 0.012	0.100 ± 0.011
Matching IoU: 0.50						
Size	**N**	**Recall (%)**	**Center Error**	**Angle (°)**	**Long Side**	**Short Side**
All	455	52.18 ± 3.34	0.055 ± 0.004	5.95 ± 0.37	0.102 ± 0.007	0.116 ± 0.010
Small	176	28.52 ± 4.78	0.069 ± 0.008	5.80 ± 0.92	0.116 ± 0.015	0.176 ± 0.024
Medium	145	51.86 ± 3.50	0.061 ± 0.005	7.19 ± 2.08	0.113 ± 0.015	0.126 ± 0.012
Large	134	83.58 ± 1.90	0.045 ± 0.002	5.50 ± 0.46	0.093 ± 0.014	0.099 ± 0.011

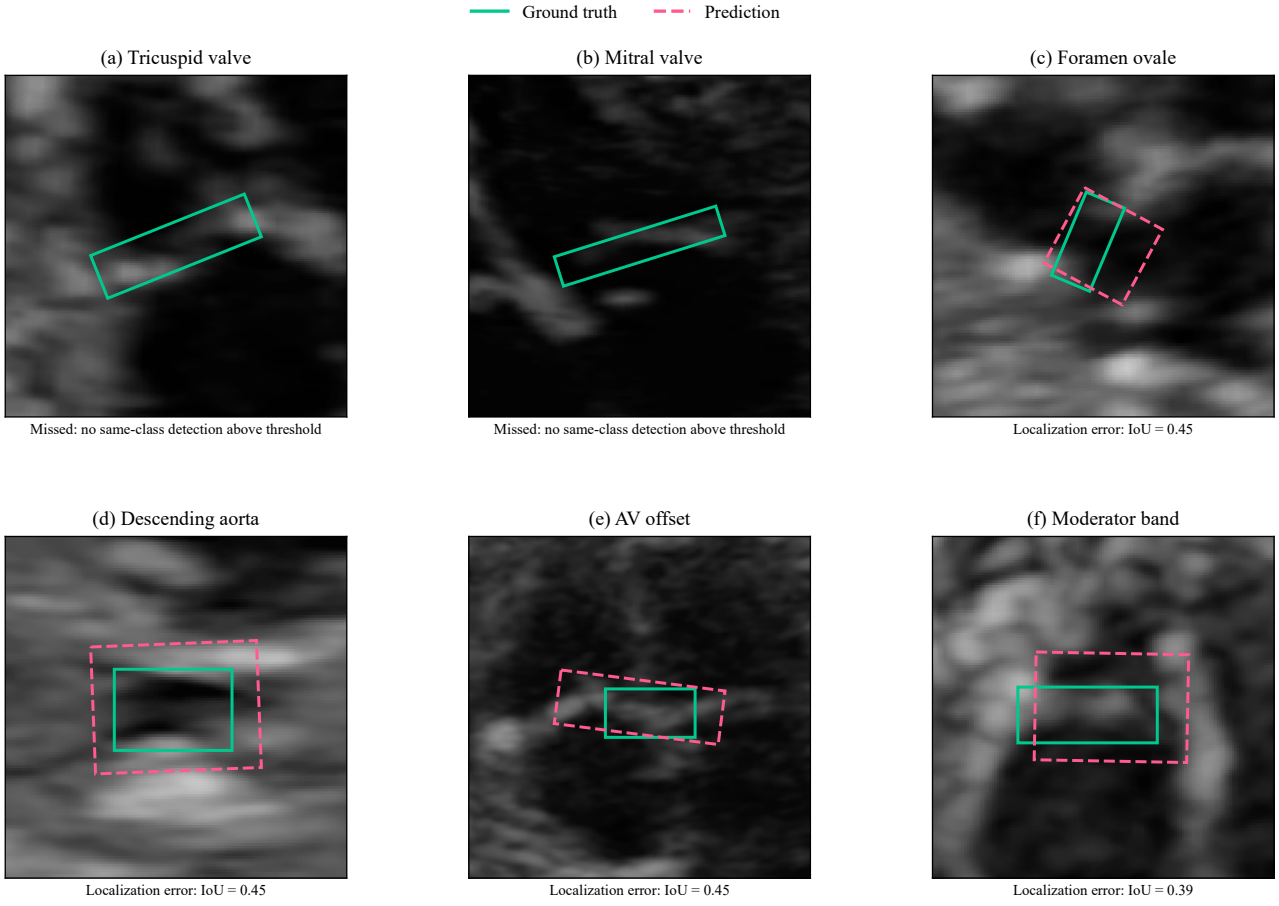


Figure 6. Representative small-structure failures on FOCUS. Enlarged regions show ground truth in solid green and same-class predictions in dashed pink. Panels (**a**,**b**) show missed targets at the validation-selected confidence threshold; panels (**c**–**f**) show localization errors with rotated IoU below 0.5. Only the target class is displayed.

4.3. Evaluation on Additional Imaging Modalities

Tables 6 and 7 report the same eight metrics for the OCD and AU-OPG datasets, with a separate model trained on each dataset. On the OCD dataset, GRIT-OBB achieved a mean AP50 of 73.14%, AP75 of 18.97%, and mAP50:95 of 30.77%, corresponding to improvements of 12.40%, 13.54%, and 11.28% over the strongest baseline values, respectively. GauCho-FCOS provided the strongest baseline AP50 and mAP50:95 values of 60.74% and 19.49%, while S^2A-Net provided the strongest baseline AP75 of 5.43%. The gains, therefore, extend to stricter cell localization.

Table 6. Summary results on the OCD dataset. The first line reports the mean ± sample SD, and the second reports the 95% training-seed t confidence interval ($n = 5$). Values in red and blue indicate the best and second-best unrounded means for each metric, respectively; rankings do not indicate statistical significance.

Model	**AP50** ↑	**AP75** ↑	**mAP50:95** ↑	**AR@100** ↑
Rotated RetinaNet	39.96 ± 14.78 [21.60, 58.31]	3.66 ± 2.04 [1.13, 6.20]	12.70 ± 5.60 [5.75, 19.66]	17.44 ± 6.08 [9.89, 24.98]
R3Det	42.45 ± 8.11 [32.37, 52.52]	1.94 ± 0.78 [0.97, 2.91]	11.85 ± 2.79 [8.38, 15.32]	16.75 ± 2.72 [13.38, 20.13]
S^2A-Net	55.97 ± 3.45 [51.69, 60.26]	5.43 ± 2.12 [2.79, 8.06]	18.36 ± 2.31 [15.49, 21.23]	23.24 ± 2.07 [20.67, 25.81]
Oriented-DETR	37.33 ± 3.36 [33.16, 41.50]	2.49 ± 0.68 [1.64, 3.34]	10.64 ± 1.45 [8.84, 12.44]	23.57 ± 0.80 [22.58, 24.55]
O2-RT-DETR	46.05 ± 5.80 [38.85, 53.25]	3.56 ± 1.52 [1.67, 5.45]	14.72 ± 2.85 [11.19, 18.26]	24.82 ± 1.82 [22.57, 27.08]
GauCho-FCOS	60.74 ± 4.36 [55.32, 66.15]	3.95 ± 1.88 [1.62, 6.27]	19.49 ± 3.41 [15.25, 23.72]	25.52 ± 2.73 [22.12, 28.91]
GRIT-OBB	73.14 ± 1.70 [71.03, 75.25]	18.97 ± 3.54 [14.58, 23.37]	30.77 ± 1.99 [28.30, 33.24]	36.61 ± 1.93 [34.21, 39.01]
Model	**Micro-F1** ↑	**Matched IoU** ↑	**FROC Mean** ↑	**oLRP$_{pool}$** ↓
Rotated RetinaNet	47.70 ± 15.56 [28.39, 67.02]	65.94 ± 2.22 [63.18, 68.70]	24.41 ± 11.32 [10.36, 38.46]	76.28 ± 7.31 [67.20, 85.35]
R3Det	51.45 ± 7.01 [42.75, 60.16]	64.63 ± 0.92 [63.49, 65.78]	23.62 ± 9.24 [12.14, 35.10]	76.06 ± 4.27 [70.76, 81.36]
S^2A-Net	63.07 ± 2.25 [60.27, 65.87]	66.84 ± 1.04 [65.55, 68.14]	35.61 ± 6.33 [27.74, 43.47]	67.97 ± 2.18 [65.26, 70.68]
Oriented-DETR	45.23 ± 2.85 [41.69, 48.78]	65.61 ± 1.04 [64.32, 66.89]	9.78 ± 2.52 [6.65, 12.90]	80.15 ± 2.02 [77.65, 82.65]
O2-RT-DETR	50.56 ± 3.84 [45.79, 55.33]	66.13 ± 1.22 [64.62, 67.64]	17.22 ± 5.48 [10.42, 24.03]	76.55 ± 3.09 [72.71, 80.39]
GauCho-FCOS	67.86 ± 2.79 [64.41, 71.32]	66.77 ± 1.46 [64.96, 68.59]	34.48 ± 7.55 [25.11, 43.85]	64.96 ± 2.96 [61.29, 68.63]
GRIT-OBB	72.95 ± 1.63 [70.92, 74.98]	72.13 ± 0.95 [70.95, 73.31]	50.57 ± 2.58 [47.37, 53.77]	56.98 ± 1.67 [54.91, 59.05]

On the AU-OPG dataset, GRIT-OBB achieved a mean AP50 of 82.65%, AP75 of 59.66%, and mAP50:95 of 50.66%, corresponding to improvements of 2.41%, 7.27%, and 3.88% over the strongest baseline values, respectively. GauCho-FCOS achieved the three strongest baseline values: 80.24%, 52.39%, and 46.78%.

Across these independently trained models, improvements in both AP50 and AP75 support the method's applicability to cell and dental localization, as well as fetal ultrasound.

Table 7. Summary results on the AU-OPG dataset. The first line reports the mean ± sample SD, and the second reports the 95% training-seed t confidence interval ($n = 5$). Values in red and blue indicate the best and second-best unrounded means for each metric, respectively; rankings do not indicate statistical significance.

Model	**AP50** ↑	**AP75** ↑	**mAP50:95** ↑	**AR@100** ↑
Rotated RetinaNet	77.01 ± 1.59 [75.04, 78.98]	47.89 ± 3.86 [43.09, 52.68]	43.73 ± 1.71 [41.60, 45.86]	60.39 ± 0.88 [59.30, 61.48]
R3Det	78.68 ± 0.87 [77.60, 79.76]	41.43 ± 3.41 [37.20, 45.66]	42.37 ± 1.16 [40.93, 43.81]	55.73 ± 1.15 [54.30, 57.16]
S^2A-Net	80.14 ± 0.43 [79.61, 80.67]	49.61 ± 2.33 [46.71, 52.51]	45.55 ± 0.89 [44.45, 46.66]	59.21 ± 0.47 [58.63, 59.80]
Oriented-DETR	72.15 ± 1.17 [70.70, 73.60]	49.02 ± 3.03 [45.26, 52.78]	43.22 ± 1.65 [41.17, 45.28]	67.83 ± 1.44 [66.04, 69.62]
O2-RT-DETR	66.62 ± 5.02 [60.38, 72.86]	41.63 ± 3.75 [36.97, 46.29]	38.51 ± 2.79 [35.05, 41.98]	61.15 ± 0.53 [60.50, 61.80]
GauCho-FCOS	80.24 ± 0.74 [79.32, 81.16]	52.39 ± 2.08 [49.81, 54.98]	46.78 ± 0.66 [45.97, 47.60]	61.83 ± 0.55 [61.15, 62.51]
GRIT-OBB	82.65 ± 1.07 [81.32, 83.97]	59.66 ± 2.97 [55.97, 63.35]	50.66 ± 1.17 [49.21, 52.12]	67.21 ± 0.62 [66.44, 67.99]
Model	**Micro-F1** ↑	**Matched IoU** ↑	**FROC Mean** ↑	**oLRP$_{pool}$** ↓
Rotated RetinaNet	78.19 ± 1.59 [76.22, 80.16]	80.47 ± 0.58 [79.75, 81.19]	61.88 ± 2.74 [58.48, 65.28]	47.60 ± 2.25 [44.80, 50.39]
R3Det	78.73 ± 1.25 [77.18, 80.28]	78.41 ± 0.72 [77.51, 79.31]	64.54 ± 2.47 [61.48, 67.61]	48.53 ± 1.34 [46.87, 50.19]
S^2A-Net	80.41 ± 0.84 [79.38, 81.45]	79.92 ± 0.39 [79.44, 80.40]	65.14 ± 2.46 [62.08, 68.20]	45.75 ± 0.99 [44.53, 46.98]
Oriented-DETR	70.78 ± 1.61 [68.78, 72.79]	81.43 ± 0.65 [80.63, 82.24]	53.22 ± 3.59 [48.77, 57.68]	54.99 ± 1.85 [52.69, 57.28]
O2-RT-DETR	67.63 ± 3.43 [63.37, 71.88]	80.04 ± 0.56 [79.34, 80.74]	48.27 ± 5.99 [40.83, 55.70]	58.14 ± 3.53 [53.76, 62.52]
GauCho-FCOS	79.74 ± 0.54 [79.07, 80.41]	80.50 ± 0.51 [79.87, 81.13]	62.71 ± 3.19 [58.75, 66.68]	46.00 ± 0.60 [45.26, 46.74]
GRIT-OBB	80.69 ± 1.85 [78.40, 82.99]	82.54 ± 0.30 [82.16, 82.91]	70.01 ± 2.61 [66.77, 73.25]	43.36 ± 2.20 [40.63, 46.09]

4.4. Performance Under Geometric and Visibility Perturbations

GRIT-OBB achieved the highest mean corrupted accuracy across the 12 perturbation conditions on all three datasets. Its r-mAP reached 11.56 ± 0.70% on FOCUS, 24.18 ± 1.10% on OCD, and 34.54 ± 0.44% on AU-OPG, compared with the strongest baseline values of 10.58%, 14.20%, and 30.32%, respectively. Figure 7 shows how this aggregate performance varies with rotation, target scale, and occlusion.

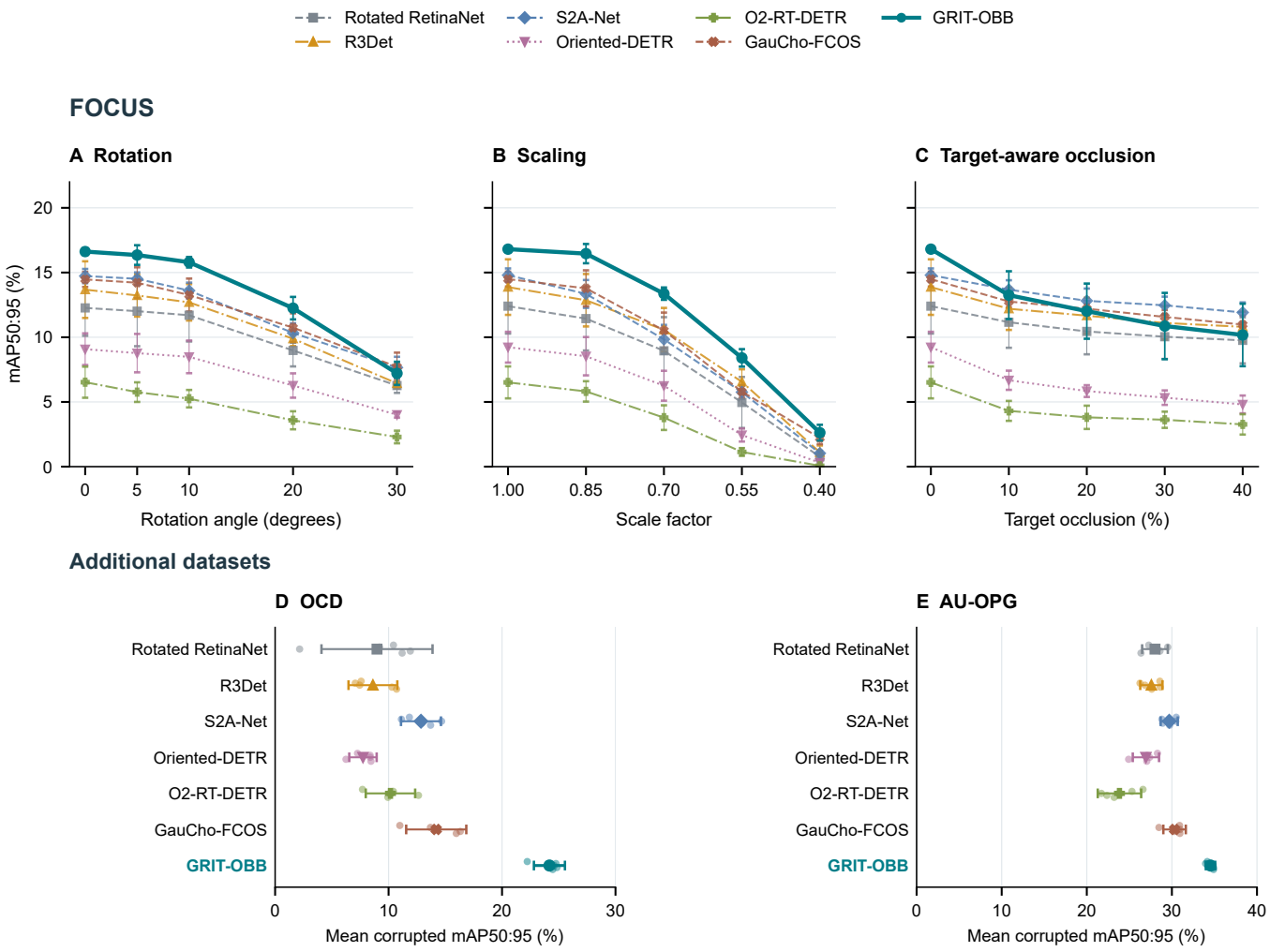


Figure 7. Detection accuracy under geometric and visibility perturbations on the FOCUS, OCD, and AU-OPG datasets. (**A**–**C**) mAP50:95 on the FOCUS dataset under rotation, downscaling, and target-aware occlusion, respectively; lines connect five-seed means, and error bars show 95% training-seed *t* intervals. (**D**,**E**) Mean accuracy across the 12 perturbation conditions (r-mAP) on the OCD and AU-OPG datasets, respectively; larger symbols and horizontal intervals indicate five-seed means and 95% training-seed *t* intervals.

On the FOCUS dataset, GRIT-OBB led at all four downscaling factors, with mAP50:95 values of 16.46%, 13.37%, 8.40%, and 2.63% as the scale factor decreased from 0.85 to 0.40. Under rotation, it obtained 16.35%, 15.80%, and 12.25% at 5°, 10°, and 20°. At 30°, its 7.21% was below S^2A-Net's 7.66% and GauCho-FCOS's 7.72%. Occlusion reduced GRIT-OBB accuracy from 13.25% at 10% coverage to 10.18% at 40%, with S^2A-Net leading at all four levels. GRIT-OBB therefore showed its clearest advantage on the FOCUS dataset under downscaling, while large rotations and occlusion remained more challenging.

Figure 8 compares GRIT-OBB with each baseline at every perturbation level. This shows whether the aggregate gains hold across individual conditions.

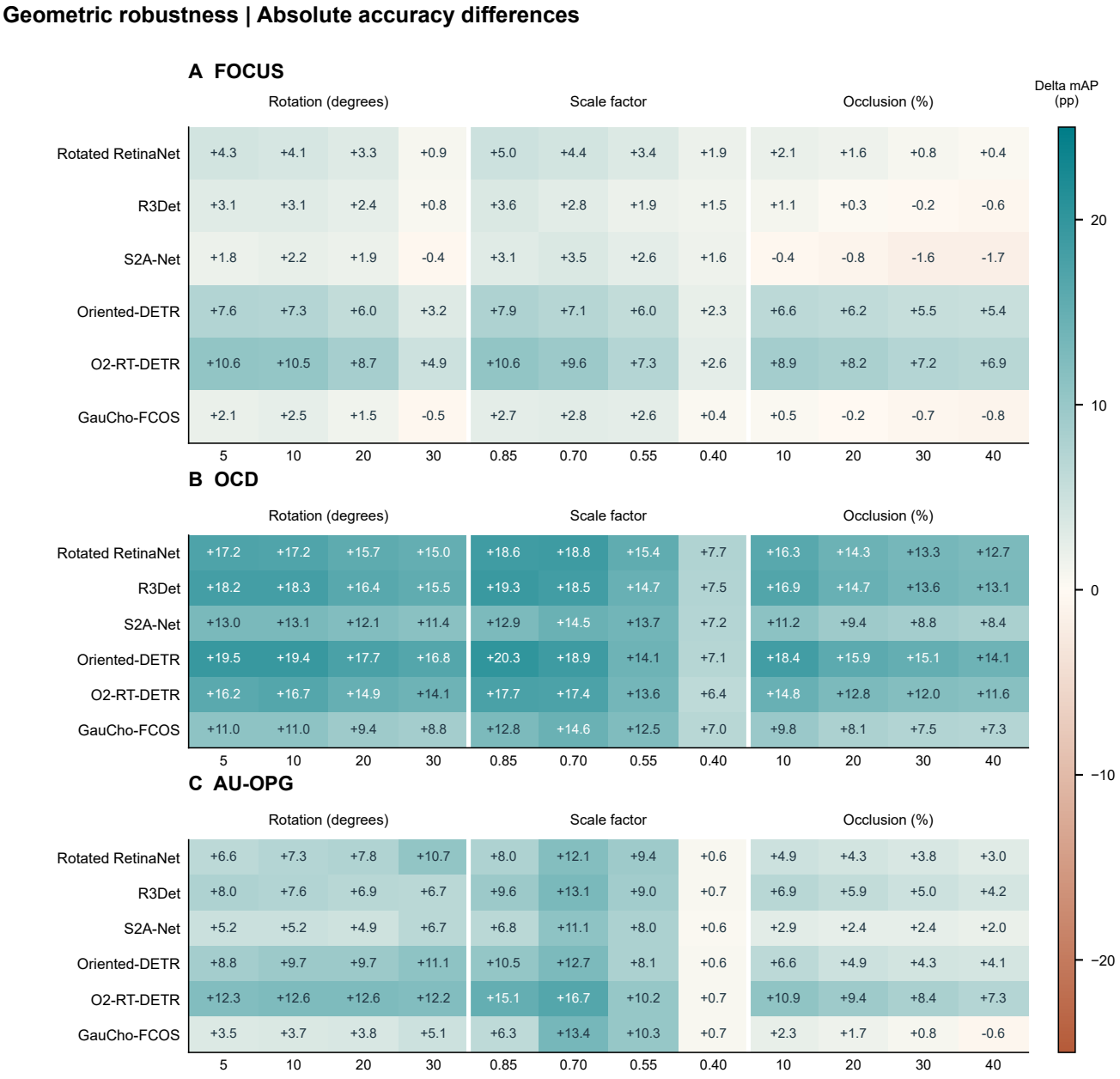


Figure 8. Condition-specific detection accuracy differences on the FOCUS, OCD, and AU-OPG datasets. Cells show the five-seed mean mAP50:95 of GRIT-OBB minus that of each baseline under 12 rotation, downscaling, and occlusion conditions. All panels share a symmetric scale centered at zero. Teal favors GRIT-OBB and brown favors the baseline; darker shades indicate larger absolute differences, and near-white cells indicate differences close to zero.

GRIT-OBB outperformed every baseline in all 72 condition-specific comparisons on OCD and in 71 of 72 comparisons on AU-OPG; the exception was GauCho-FCOS at 40% occlusion. On FOCUS, all downscaling comparisons favored GRIT-OBB, while negative differences occurred under the strongest rotation and several occlusion conditions. At the smallest AU-OPG scale, GRIT-OBB reached 0.71% mAP50:95, showing the substantial loss of accuracy when target resolution is severely reduced.

Tables 8–10 distinguish absolute accuracy from retention relative to clean performance. GRIT-OBB led r-mAP and worst-case severe accuracy on all three datasets. It also achieved the highest mean RR-AUC on OCD and AU-OPG, at 84.01% and 74.67%, respectively. On

FOCUS, its RR-AUC was 75.13%, compared with 77.75% for R3Det. Thus, the FOCUS advantage is strongest in absolute detection accuracy, whereas OCD and AU-OPG show gains in both accuracy and relative retention.

Table 8. Robustness results on the FOCUS dataset. Metrics summarize clean accuracy, mean accuracy across 12 perturbation conditions (r-mAP), clean-normalized accuracy retention (RR-AUC), and worst-case accuracy across the most severe conditions. Entries report the mean ± sample SD, followed by the 95% training-seed t interval ($n = 5$). Red and blue indicate the best and second-best means, respectively, without implying statistical significance.

Model	Clean mAP ↑	r-mAP ↑	RR-AUC ↑	Worst-Case Severe Accuracy ↑
Rotated RetinaNet	12.40 ± 1.71 [10.28, 14.52]	8.87 ± 1.40 [7.14, 10.61]	77.18 ± 1.60 [75.19, 79.16]	0.76 ± 0.37 [0.30, 1.22]
R3Det	13.87 ± 1.73 [11.73, 16.02]	9.93 ± 0.90 [8.81, 11.04]	77.75 ± 3.14 [73.85, 81.65]	1.11 ± 0.42 [0.58, 1.63]
S^2A-Net	14.80 ± 0.42 [14.28, 15.32]	10.58 ± 0.40 [10.09, 11.08]	76.85 ± 1.74 [74.69, 79.00]	1.03 ± 0.14 [0.85, 1.21]
Oriented-DETR	9.23 ± 0.96 [8.04, 10.42]	5.64 ± 0.49 [5.04, 6.25]	68.46 ± 1.45 [66.65, 70.26]	0.31 ± 0.15 [0.11, 0.50]
O2-RT-DETR	6.51 ± 0.99 [5.28, 7.75]	3.56 ± 0.43 [3.02, 4.10]	61.89 ± 2.68 [58.57, 65.22]	0.07 ± 0.04 [0.01, 0.12]
GauCho-FCOS	14.49 ± 0.46 [13.92, 15.07]	10.48 ± 0.49 [9.87, 11.09]	77.51 ± 1.40 [75.77, 79.25]	2.21 ± 0.38 [1.74, 2.69]
GRIT-OBB	16.81 ± 0.19 [16.57, 17.04]	11.56 ± 0.70 [10.70, 12.43]	75.13 ± 3.11 [71.26, 78.99]	2.63 ± 0.49 [2.02, 3.23]

Table 9. Robustness results on the OCD dataset. Clean mAP denotes mAP50:95 on unperturbed images with shared preprocessing. The r-mAP averages mAP50:95 across four rotation angles (5°, 10°, 20°, 30°), four scale factors (0.85, 0.70, 0.55, 0.40), and four occlusion fractions (0.10, 0.20, 0.30, 0.40). RR-AUC is the mean area under the clean-normalized accuracy curve across the three perturbation families. Worst-case severe accuracy is the lowest mAP50:95 under 30° rotation, 0.40 scaling, or 40% occlusion. The first line reports the mean ± sample SD, and the second reports the 95% training-seed t confidence interval ($n = 5$). Values in red and blue indicate the best and second-best unrounded means for each metric, respectively; rankings do not indicate statistical significance.

Model	Clean mAP ↑	r-mAP ↑	RR-AUC ↑	Worst-Case Severe Accuracy ↑
Rotated RetinaNet	12.70 ± 5.60 [5.75, 19.65]	8.98 ± 3.94 [4.09, 13.87]	77.33 ± 1.43 [75.56, 79.11]	1.34 ± 0.80 [0.35, 2.33]
R3Det	11.85 ± 2.79 [8.38, 15.32]	8.62 ± 1.73 [6.47, 10.77]	79.71 ± 5.66 [72.68, 86.73]	1.53 ± 0.48 [0.94, 2.12]
S^2A-Net	18.36 ± 2.31 [15.49, 21.23]	12.85 ± 1.42 [11.09, 14.62]	76.71 ± 1.88 [74.38, 79.05]	1.82 ± 0.40 [1.32, 2.31]
Oriented-DETR	10.68 ± 1.32 [9.05, 12.32]	7.74 ± 0.97 [6.53, 8.95]	78.69 ± 2.56 [75.52, 81.87]	1.95 ± 0.65 [1.15, 2.75]
O2-RT-DETR	14.72 ± 2.85 [11.19, 18.26]	10.17 ± 1.75 [7.99, 12.34]	75.33 ± 3.21 [71.34, 79.32]	2.63 ± 0.88 [1.53, 3.72]
GauCho-FCOS	19.77 ± 2.62 [16.53, 23.02]	14.20 ± 2.13 [11.55, 16.85]	77.55 ± 3.35 [73.40, 81.71]	2.03 ± 0.94 [0.87, 3.19]
GRIT-OBB	30.77 ± 1.99 [28.30, 33.24]	24.18 ± 1.10 [22.81, 25.54]	84.01 ± 1.92 [81.63, 86.38]	9.06 ± 0.93 [7.90, 10.21]

Table 10. Robustness results on the AU-OPG dataset. Clean mAP denotes mAP50:95 on unperturbed images with shared preprocessing. The r-mAP averages mAP50:95 across four rotation angles (5°, 10°, 20°, 30°), four scale factors (0.85, 0.70, 0.55, 0.40), and four occlusion fractions (0.10, 0.20, 0.30, 0.40). RR-AUC is the mean area under the clean-normalized accuracy curve across the three perturbation families. Worst-case severe accuracy is the lowest mAP50:95 under 30° rotation, 0.40 scaling, or 40% occlusion. The first line reports the mean ± sample SD, and the second reports the 95% training-seed *t* confidence interval ($n = 5$). Values in red and blue indicate the best and second-best unrounded means for each metric, respectively; rankings do not indicate statistical significance.

Model	Clean mAP ↑	r-mAP ↑	RR-AUC ↑	Worst-Case Severe Accuracy ↑
Rotated RetinaNet	43.73 ± 1.71 [41.60, 45.86]	28.01 ± 1.22 [26.50, 29.52]	70.53 ± 0.51 [69.90, 71.17]	0.08 ± 0.06 [0.01, 0.15]
R3Det	42.37 ± 1.16 [40.93, 43.81]	27.59 ± 1.05 [26.28, 28.89]	71.43 ± 1.01 [70.17, 72.69]	0.05 ± 0.03 [0.01, 0.08]
S^2A-Net	45.55 ± 0.89 [44.45, 46.66]	29.69 ± 0.81 [28.69, 30.69]	71.54 ± 0.62 [70.76, 72.31]	0.14 ± 0.08 [0.05, 0.24]
Oriented-DETR	43.33 ± 1.75 [41.15, 45.50]	26.95 ± 1.24 [25.40, 28.49]	68.82 ± 0.88 [67.73, 69.90]	0.11 ± 0.05 [0.04, 0.17]
O2-RT-DETR	38.51 ± 2.79 [35.05, 41.97]	23.84 ± 2.05 [21.29, 26.39]	68.35 ± 1.52 [66.46, 70.24]	0.05 ± 0.05 [−0.01, 0.10]
GauCho-FCOS	46.98 ± 0.62 [46.21, 47.75]	30.32 ± 1.06 [28.99, 31.64]	70.91 ± 1.68 [68.83, 73.00]	0.05 ± 0.05 [−0.02, 0.11]
GRIT-OBB	50.66 ± 1.17 [49.21, 52.12]	34.54 ± 0.44 [33.99, 35.09]	74.67 ± 1.52 [72.77, 76.56]	0.71 ± 0.19 [0.47, 0.95]

Across the three datasets, GRIT-OBB achieved the highest average detection accuracy under the evaluated perturbations, with paired bootstrap comparisons supporting the r-mAP gains over all six baselines (all 18 Holm-adjusted $p \leq 0.0026$). For the following comparisons, differences were calculated by subtracting the baseline r-mAP from the GRIT-OBB r-mAP, with both scores expressed as percentages. The mean differences were 0.98 on the FOCUS dataset (S^2A-Net; 95% paired bootstrap confidence interval [0.37, 1.72]; Holm-adjusted $p = 0.0026$), 9.98 on the OCD dataset (GauCho-FCOS; 95% interval [4.90, 13.91]), and 4.22 on the AU-OPG dataset (GauCho-FCOS; 95% interval [3.55, 4.99]). Its advantage is consistent across downscaling conditions in FOCUS and across nearly all individual conditions in OCD and AU-OPG. The mean RR-AUC difference relative to the strongest baseline, calculated from scores expressed as percentages, was larger on the OCD dataset (4.30) than on the AU-OPG dataset (3.13), indicating that robustness varies with both image modality and perturbation type.

4.5. Component Ablation

The final experiment tested the contributions of geometric correction, regional interaction, and joint overlap supervision on the FOCUS dataset. The three ablated configurations are labeled w/o geometry correction, w/o geometry relation, and w/o dual-stage IoU, respectively, removing the correction pathway, the relation residual, or both stage-specific IoU losses. Without geometric correction, horizontal proposals became the second-stage references and the first-stage regression and IoU losses were omitted; second-stage sampling retained the same context factors. The relation and IoU ablations retained the correction pathway. Figure 9 compares AP across overlap thresholds and sensitivity across false-positive budgets. GRIT-OBB achieved the highest mean AP50, AP75, and mAP50:95; Table 11 reports all eight metrics and training-run variability.

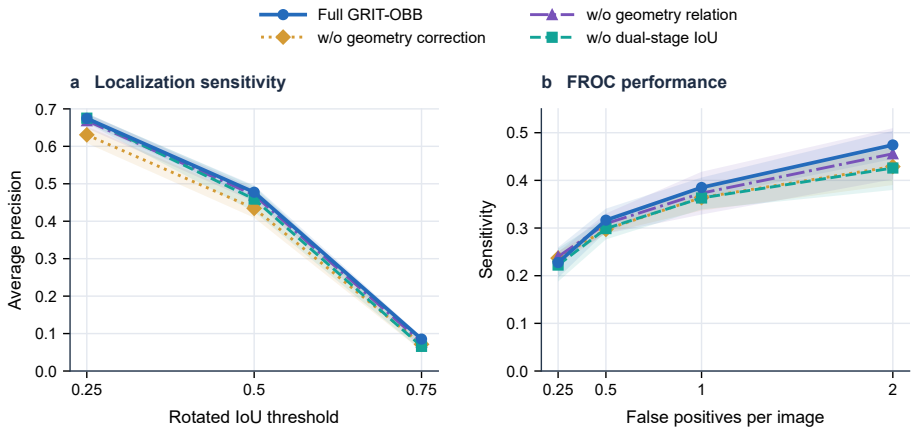


Figure 9. Component ablation on the FOCUS dataset. (**a**) Rotated AP at IoU thresholds 0.25, 0.50, and 0.75. (**b**) FROC sensitivity at 0.25, 0.5, 1, and 2 false positives per image. Markers show five-seed means, with pointwise 95% training-seed t intervals as shaded bands. Straight lines connect the mean values at the evaluated IoU thresholds and false-positive budgets. FROC budgets were obtained by scanning prediction scores, independently of the validation-selected operating threshold. The ablated configurations remove geometric correction (w/o geometry correction), the geometric relation residual (w/o geometry relation), or both stage-specific IoU losses (w/o dual-stage IoU). All configurations used the same fixed test partition.

Table 11. Component ablation results on the FOCUS dataset. The ablated configurations remove geometric correction (w/o geometry correction), the geometric relation residual (w/o geometry relation), or both stage-specific IoU losses (w/o dual-stage IoU). The first line reports the mean ± sample SD, and the second reports the 95% training-seed t confidence interval ($n = 5$). Values in red and blue indicate the best and second-best unrounded means for each metric, respectively; rankings do not indicate statistical significance.

Model	AP50 ↑	AP75 ↑	mAP50:95 ↑	AR@100 ↑
GRIT-OBB	47.74 ± 1.46 [45.93, 49.56]	8.53 ± 0.59 [7.80, 9.25]	16.81 ± 0.19 [16.57, 17.04]	30.28 ± 0.77 [29.32, 31.24]
w/o geometry correction	43.44 ± 2.03 [40.92, 45.96]	7.13 ± 1.37 [5.44, 8.83]	15.20 ± 0.90 [14.09, 16.32]	27.70 ± 1.02 [26.43, 28.96]
w/o geometry relation	46.99 ± 1.39 [45.26, 48.72]	7.68 ± 1.12 [6.30, 9.07]	16.73 ± 0.58 [16.02, 17.45]	29.39 ± 0.62 [28.62, 30.16]
w/o dual-stage IoU	45.91 ± 2.75 [42.49, 49.32]	6.59 ± 0.90 [5.47, 7.70]	15.78 ± 0.92 [14.64, 16.93]	28.87 ± 0.90 [27.76, 29.98]
Model	**Micro-F1 ↑**	**Matched IoU ↑**	**FROC Mean ↑**	**oLRP$_{pool}$ ↓**
GRIT-OBB	56.94 ± 2.67 [53.62, 60.25]	69.11 ± 0.76 [68.17, 70.05]	35.09 ± 1.97 [32.64, 37.53]	71.43 ± 1.10 [70.07, 72.79]
w/o geometry correction	52.30 ± 2.78 [48.84, 55.75]	68.75 ± 0.49 [68.14, 69.36]	33.18 ± 1.22 [31.66, 34.69]	74.07 ± 1.66 [72.02, 76.13]
w/o geometry relation	55.90 ± 2.60 [52.68, 59.12]	69.25 ± 0.60 [68.51, 69.99]	34.45 ± 2.36 [31.52, 37.38]	72.13 ± 2.08 [69.54, 74.71]
w/o dual-stage IoU	51.82 ± 3.76 [47.15, 56.49]	68.60 ± 0.94 [67.44, 69.76]	32.76 ± 2.39 [29.78, 35.73]	74.15 ± 1.85 [71.86, 76.44]

Removing geometric correction reduced AP50 from 47.74% to 43.44%, mAP50:95 from 16.81% to 15.20%, and AR@100 from 30.28% to 27.70%. The concurrent decreases in detection and recall support the role of corrected geometry in establishing the second-stage sampling frame. Joint removal of the two IoU losses reduced AP75 from 8.53% to 6.59% and mAP50:95 from 16.81% to 15.78%. At IoU 0.25, the two configurations were close: 67.52% without IoU supervision and 67.46% for GRIT-OBB. As the overlap criterion became

stricter, direct supervision produced a clearer advantage. Micro-F1 also decreased from 56.94% to 51.82% after removal. This contrast links the combined two-stage objective to localization quality beyond coarse target retrieval. The geometry relation residual was associated with a small increase in mean AP50, from 46.99% for w/o geometry relation to 47.74% for GRIT-OBB, and in AP75, from 7.68% to 8.53%. FROC mean sensitivity increased from 34.45% to 35.09%. These differences are within the observed training-seed variability, with overlapping 95% confidence intervals. The mAP50:95 values were 16.73% for w/o geometry relation and 16.81% for GRIT-OBB. Mean matched IoU was also similar, at 69.25% and 69.11%, respectively, so these results did not show a clear effect of the relation residual on the overlap quality of matched detections. The FROC curves show how the configurations behave under different false-positive budgets. At 0.25 false positives per image, w/o geometry relation and w/o geometry correction reached sensitivities of 23.99% and 23.69%, respectively, compared with 22.76% for GRIT-OBB. GRIT-OBB led at the other three budgets, reaching 31.65%, 38.51%, and 47.43% at 0.5, 1, and 2 false positives per image, respectively. Its four-budget mean was 35.09%. The complete method therefore achieved the highest average sensitivity across the evaluated budgets, although removing either the geometry relation or geometry correction yielded slightly higher sensitivity at 0.25 false positives per image.

The ablation results support the roles of geometric correction, regional interaction, and direct overlap supervision in GRIT-OBB. Corrected proposals provide a sampling reference for improved detection and recall; regional interaction is associated with small increases in detection metrics within the observed training-seed variability; and overlap supervision improves localization at stricter IoU thresholds. Their integration yielded the highest mean AP50, AP75, mAP50:95, and FROC sensitivity among the evaluated configurations, supporting the use of proposal geometry to connect regional representation with successive localization stages.

4.5.1. Ablation of the Geometry-Guided Relation Residual

We separately ablated the three components of the geometry-guided relation residual. Removing positional encoding sets the Gaussian encoding added to the query and key vectors to zero, while retaining the Wasserstein attention bias. Removing the Wasserstein bias sets the geometric attention bias term to zero while retaining positional encoding. Removing residual gating replaces the learned scalar gate with a fixed value of one, so the relation update is added directly to the regional feature. The remaining relation components are retained in each variant. Each variant was trained with seeds 0-4 on the fixed FOCUS partition, with checkpoints selected by validation AP50. The full-model comparison uses the corresponding five existing training runs.

As shown in Table 12, removing positional encoding produced the largest reduction in mean AP50, from 47.74% to 44.47%, and reduced AP75 from 8.53% to 6.82%. Removing the Wasserstein bias reduced these metrics to 45.56% and 7.14%, respectively. Both variants also had lower mean recall, Micro-F1, matched IoU, and FROC sensitivity, and higher pooled optimal LRP than the full model. These results support the contribution of both forms of geometric information to regional interaction.

The ungated variant was closer to the full model in mAP50:95 (16.78% versus 16.81%), but had lower mean AP50 (46.53% versus 47.74%) and Micro-F1 (54.45% versus 56.94%). Its mean matched IoU was slightly higher (69.43% versus 69.11%). The full model achieved higher AP50 and Micro-F1 than the ungated variant, while mAP50:95 remained nearly unchanged (16.81% versus 16.78%). The ungated variant achieved a slightly higher matched IoU (69.43% versus 69.11%).

Table 12. Internal ablation of the geometry-guided relation residual on FOCUS. Values are mean and sample SD across five training seeds, expressed as percentages. Higher values are better, except for pooled optimal LRP, where lower values are better. Red and blue indicate the best and second-best unrounded means for each metric, respectively; rankings do not indicate statistical significance.

Variant	AP50	AP75	mAP50:95	AR@100
Full GRIT-OBB	47.74 ± 1.46	8.53 ± 0.59	16.81 ± 0.19	30.28 ± 0.77
w/o positional encoding	44.47 ± 1.74	6.82 ± 0.67	15.51 ± 0.75	28.79 ± 1.01
w/o Wasserstein bias	45.56 ± 1.25	7.14 ± 1.22	16.02 ± 0.82	29.16 ± 0.51
w/o residual gating	46.53 ± 1.86	7.87 ± 1.31	16.78 ± 0.78	29.94 ± 0.82
Variant	**Micro-F1**	**Matched IoU**	**FROC Mean**	**oLRP$_{pool}$**
Full GRIT-OBB	56.94 ± 2.67	69.11 ± 0.76	35.09 ± 1.97	71.43 ± 1.10
w/o positional encoding	53.24 ± 1.13	68.63 ± 0.64	33.18 ± 1.73	74.17 ± 0.78
w/o Wasserstein bias	52.46 ± 3.69	68.53 ± 0.55	33.25 ± 2.43	73.93 ± 1.31
w/o residual gating	54.45 ± 2.64	69.43 ± 0.99	34.23 ± 1.35	72.70 ± 1.90

4.5.2. Sensitivity to Rotated IoU Loss Weights

We varied the proposal-correction and final-refinement IoU weights individually around the default pair (2,2), halving or doubling one coefficient while keeping the other fixed. All configurations used training seed 0 and the fixed FOCUS partition. The four additional runs used FP32, batch size 1, a maximum of 50 epochs, and validation-AP50 checkpoint selection. Context enlargement remained fixed at (1.2,1.4). The default row in Table 13 uses the existing full-model seed-0 run.

Table 13. Sensitivity to the two rotated IoU loss weights on FOCUS. Each row reports one seed-0 run; metric values are percentages. Higher values are better except for the pooled optimal LRP. Red and blue indicate the best and second-best values for each metric, respectively.

λ_1, λ_2	AP50	AP75	mAP50:95	AR@100
(2, 2) (default)	47.67	9.01	16.84	29.78
(1, 2)	45.97	6.64	15.72	28.95
(4, 2)	47.29	7.39	16.88	29.96
(2, 1)	47.31	9.51	17.55	29.52
(2, 4)	44.73	5.70	14.97	28.92
λ_1, λ_2	**Micro-F1**	**Matched IoU**	**FROC Mean**	**oLRP$_{pool}$**
(2, 2) (default)	58.24	69.61	36.84	71.03
(1, 2)	53.89	67.90	33.15	73.00
(4, 2)	53.51	69.42	33.24	72.84
(2, 1)	57.31	69.81	35.23	70.95
(2, 4)	51.71	68.64	31.04	75.14

The default weights (2,2) achieved the highest AP50 (47.67%), Micro-F1 (58.24%), and FROC mean sensitivity (36.84%). Reducing the final-refinement weight to 1 increased AP75 from 9.01% to 9.51%, mAP50:95 from 16.84% to 17.55%, and matched IoU from 69.61% to 69.81%, while reducing pooled optimal LRP from 71.03% to 70.95%. Increasing this weight to 4 worsened all eight metrics relative to the default. Increasing the proposal-correction weight to 4 yielded the highest AR@100 (29.96%), whereas reducing it to 1 worsened all eight metrics. These results show that stronger overlap weighting does not consistently improve localization. The equal-weight default strikes a balance between detection performance and localization accuracy, whereas a lower final-stage weight favors stricter overlap metrics in this comparison.

4.5.3. Sensitivity to Context Enlargement

We compared the default context enlargement factors (1.2,1.4) with no enlargement (1.0,1.0), smaller factors (1.1,1.2), larger factors (1.3,1.6), and reversed factors (1.4,1.2). These factors scale the two local side lengths of the oriented sampling region. All configurations used training seed 0 and the fixed FOCUS partition, with both IoU loss weights fixed at 2. The four additional runs used FP32, batch size 1, a maximum of 50 epochs, and validation-AP50 checkpoint selection; the default row in Table 14 uses the existing full-model seed-0 run.

Table 14. Sensitivity to context enlargement on FOCUS. Each row reports one seed-0 run; values are percentages. Higher values are better except for the pooled optimal LRP. Red and blue indicate the best and second-best values, respectively.

α_w, α_h	AP50	AP75	mAP50:95	AR@100
(1.2, 1.4) (default)	47.67	9.01	16.84	29.78
(1.0, 1.0)	44.27	6.96	15.74	29.32
(1.1, 1.2)	45.54	7.77	16.39	30.46
(1.3, 1.6)	43.68	8.55	15.93	29.43
(1.4, 1.2)	45.90	5.50	16.01	28.57
α_w, α_h	Micro-F1	Matched IoU	FROC Mean	$oLRP_{pool}$
(1.2, 1.4) (default)	58.24	69.61	36.84	71.03
(1.0, 1.0)	54.71	69.87	30.87	73.44
(1.1, 1.2)	54.77	69.74	32.36	73.02
(1.3, 1.6)	53.45	69.88	31.85	73.52
(1.4, 1.2)	50.61	68.21	31.61	74.10

The default factors achieved the highest AP50 (47.67%), AP75 (9.01%), mAP50:95 (16.84%), Micro-F1 (58.24%), and FROC mean sensitivity (36.84%), together with the lowest pooled optimal LRP (71.03%). Smaller factors (1.1,1.2) yielded the highest AR@100 (30.46%), whereas larger factors (1.3,1.6) yielded the highest matched IoU (69.88%). Reversing the default factors reduced AP75 from 9.01% to 5.50%, indicating that the allocation of context between the two local directions affects localization.

To examine dependence on target geometry, we grouped ground truth by normalized area and long-to-short-side aspect ratio using training-set tertiles. The area boundaries were 0.00514 and 0.01394, and the aspect-ratio boundaries were 1.26923 and 1.90792. Group membership was determined by ground-truth geometry. We performed confidence-ordered, class-aware one-to-one matching at IoU 0.5 over all saved post-processing candidates, without an additional operating-threshold filter. Each group in Table 15 reports ground-truth recall and mean IoU among matched detections; these summaries therefore differ from the operating-point metrics in the overall table.

Increasing the factors to (1.3,1.6) raised recall for small targets from 66.48% to 73.86% and for high-aspect-ratio targets from 71.66% to 75.94%, but reduced large-target recall from 97.01% to 88.81% and overall AP50 from 47.67% to 43.68%. The default factors yielded the highest recall and matched IoU for large targets, at 97.01% and 72.18%, respectively. Thus, a larger sampling region can improve the retrieval of small or elongated targets while reducing performance elsewhere. The default factors provide the strongest overall detection performance among the tested configurations.

Table 15. Context sensitivity by target size and aspect ratio. Each pair of columns reports recall at IoU 0.5 (R) and mean matched IoU (I), in percent. Size groups contain 176, 145, and 134 targets; aspect-ratio groups contain 135, 133, and 187 targets. Higher values are better. Red and blue indicate the best and second-best values within each column, with ties sharing a color.

α_w, α_h	Small R	Small I	Medium R	Medium I	Large R	Large I
(1.2, 1.4) (default)	66.48	61.23	77.24	65.52	97.01	72.18
(1.0, 1.0)	67.05	61.17	79.31	65.09	92.54	71.55
(1.1, 1.2)	71.59	61.22	80.00	65.03	94.03	71.76
(1.3, 1.6)	73.86	61.26	79.31	65.09	88.81	71.96
(1.4, 1.2)	56.82	61.34	80.00	65.82	95.52	69.81
α_w, α_h	**Low AR R**	**Low AR I**	**Mid AR R**	**Mid AR I**	**High AR R**	**High AR I**
(1.2, 1.4) (default)	82.22	67.65	85.71	70.21	71.66	62.48
(1.0, 1.0)	86.67	67.20	87.97	68.03	65.78	63.05
(1.1, 1.2)	89.63	66.97	86.47	70.12	70.59	61.60
(1.3, 1.6)	82.96	66.97	82.71	69.30	75.94	62.59
(1.4, 1.2)	80.74	67.98	87.97	67.29	63.10	62.90

4.6. Computational Efficiency

We evaluated parameter counts, computational cost, and preprocessing time. All methods used 1024 × 1024 inputs, batch size 1, and FP32 on an NVIDIA RTX 4090 D, with TF32 disabled and one CPU tensor-operation thread. Each method retained its compatible software environment.

Parameter counts include frozen parameters. FLOP count convolution, matrix operations, and custom sampling operators, using two operations per multiply–accumulate; scalar geometry, nonlinearities, normalization, softmax, indexing, and NMS are excluded. Preprocessing time included data loading, input preparation, and host-to-device transfer and was summarized as the mean ± sample standard deviation across 50 test images from the FOCUS dataset. Normalization implemented within the detector was excluded from this measurement.

Table 16 shows that GRIT-OBB has the fewest parameters and the third-lowest FLOPs among the compared methods. Its preprocessing time was 15.66 ± 1.40 ms per image. These results characterize model size, counted arithmetic operations, and the measured preprocessing workload; preprocessing times also depend on where each implementation performs input preparation.

We additionally benchmarked end-to-end latency and GPU memory using the existing seed-0 checkpoints on the same FOCUS test images for all seven methods. Each method ran sequentially in its compatible environment, with the hardware, precision, input resolution, and batch size settings specified above. After 20 warm-up calls, we timed one pass over the test images with GPU synchronization. Timing included file loading and decoding, input preparation, host-to-device transfer, native prediction with normalization and post-processing, and output recovery to CPU coordinates. Model initialization and result serialization were excluded. Latency is reported as the mean and sample SD across images, together with the 95th percentile (P95). Memory values are the largest per-image PyTorch-allocated and reserved GPU memory.

Table 16. Parameter counts, computational cost, and preprocessing time on the FOCUS dataset at 1024 × 1024 input resolution. Preprocessing time includes data loading, input preparation, and host-to-device transfer and is reported as mean ± sample SD across 50 test images. Red and blue indicate the lowest and second-lowest values in each column, respectively.

Method	Parameters (M) ↓	GFLOPs ↓	Preprocess (ms) ↓
Rotated RetinaNet	36.64	430.26	2.88 ± 1.97
R3Det	47.62	909.22	3.45 ± 2.12
S^2A-Net	36.47	445.49	3.29 ± 1.88
Oriented-DETR	57.08	849.94	21.85 ± 1.33
O2-RT-DETR	42.81	339.53	8.92 ± 1.11
GauCho-FCOS	32.15	412.16	32.79 ± 1.65
GRIT-OBB	28.58	418.41	15.66 ± 1.40

Table 17 shows that GRIT-OBB required 112.28 ms per image on average, with a P95 latency of 149.21 ms. It was faster than Oriented-DETR (244.67 ms) but slower than the other five baselines (39.40-80.49 ms). Its native prediction stage took 99.23 ms, while input preparation and transfer took 12.86 ms and output recovery took 0.19 ms on average. Peak allocated and reserved GPU memory were 547.57 and 770.00 MiB, respectively.

Table 17. End-to-end inference latency and GPU memory on FOCUS test images. Latency is in milliseconds per image; memory is in MiB. Lower values are better. Red and blue indicate the lowest and second-lowest values in each column, respectively.

Method	Latency (Mean ± SD)	P95	Peak Allocated	Peak Reserved
Rotated RetinaNet	48.27 ± 2.03	51.24	363.01	456.00
R3Det	80.49 ± 1.28	82.35	406.20	504.00
S^2A-Net	52.11 ± 1.31	54.16	370.47	594.00
Oriented-DETR	244.67 ± 4.02	251.11	785.23	1090.00
O2-RT-DETR	39.40 ± 1.03	41.02	396.29	482.00
GauCho-FCOS	61.39 ± 1.86	63.73	345.87	360.00
GRIT-OBB	112.28 ± 17.20	149.21	547.57	770.00

5. Discussion

5.1. Geometry as a Shared Reference for Detection

The results support the use of proposal geometry as a shared reference for regional feature extraction and subsequent localization. GRIT-OBB corrects proposals before feature sampling and supervises both the corrected reference and the final detection. The reductions in detection and recall after removing geometric correction, together with the decline in AP75 after removing overlap supervision, suggest that proposal correction and overlap supervision contribute to different aspects of detection and localization. Each stage receives direct supervision, while shared feature learning connects the two stages.

The class-level results clarify where this design is most useful. Four methods achieved 100% AP50 for whole-heart detection, while GRIT-OBB's largest gains occurred for the interventricular septum, mitral valve, and right ventricle. These improvements are consistent with the value of aligning regional features to structures embedded within a shared anatomical field. Regional interaction further supports detection by exchanging contextual information between aligned regions. Its geometric bias describes spatial extent and proximity, allowing contextual features to complement target-aligned detail without imposing a fixed anatomical topology.

5.2. Robustness and Applicability Across Modalities

The perturbation experiments show how GRIT-OBB performance varies with target orientation, resolution, and visibility. GRIT-OBB maintained a consistent advantage under downscaling on FOCUS, while large rotations and occlusion produced less uniform gains. This pattern suggests that aligned sampling supports detection when spatial detail is reduced, while occlusion removes target content that feature alignment alone cannot recover. Detection accuracy under perturbations and relative retention capture these distinctions: GRIT-OBB had the highest r-mAP across all three datasets, but the highest relative retention only on OCD and AU-OPG.

The results on the OCD microscopy dataset and the AU-OPG panoramic dental radiography dataset extend the evidence to different image appearances and annotation geometries. Improvements in strict localization across both datasets suggest that the design that combines proposal correction, feature alignment, and bounding-box refinement is applicable beyond fetal ultrasound. These results suggest that the method is applicable across imaging modalities; evaluation across institutions and acquisition systems is the next step toward assessing broader generalization.

5.3. Limitations and Future Directions

The evaluation used small fixed partitions, including 50 FOCUS test images. Five training runs characterize the variability in optimization within these partitions, whereas independent cohorts are needed to assess population-level generalization.

Each image was annotated by a single expert, so inter-observer agreement could not be quantified; annotation variability, especially for small structures, may therefore contribute to the reported errors. Future work should quantify annotation variability through multi-reader annotation.

The size-stratified error analysis helps explain the gap between coarse detection and precise localization. Small targets combined lower recall with greater relative center displacement and short-side error than large targets. For a small structure, a given pixel displacement occupies a larger fraction of its extent, making overlap particularly sensitive to center and boundary errors. Improving fine spatial detail and short-axis dimension estimation is therefore a priority for stricter localization. The angular-error summaries describe matched detections, so they do not rule out large orientation errors among unmatched targets.

Further analysis is needed to resolve the mechanisms underlying the observed gains. The joint overlap ablation removed both stage-specific terms, so their individual effects cannot be separated in this comparison. Direct measurements of proposal error and sampling alignment, together with targeted analysis of context extent and regional interactions, could help explain the remaining difficulties in strict localization and in classes such as the cardiac apex, valve-offset landmarks, and the moderator band.

GRIT-OBB had the fewest parameters among the compared methods, at 28.58 million, and the third-lowest computational cost, at 418.41 GFLOPs. The mean preprocessing time was 15.66 ms per image. Its measured end-to-end latency was 112.28 ms per image, corresponding to approximately 8.9 images per second at batch size 1, with peak allocated GPU memory of 547.57 MiB. Although faster than Oriented-DETR, it remained slower than the other five baselines. Further optimization and evaluation within clinical workflows are needed to establish its suitability for live ultrasound assistance. The small-structure failure cases also indicate that predictions require clinical review, particularly for missed valves and imprecisely localized intracardiac landmarks.

Generalization across ultrasound scanners, gestational ages, imaging conditions, and clinical populations remains to be evaluated. Differences in scanner characteristics and

acquisition settings may affect image appearance, while gestational age changes the size and relative proportions of cardiac structures. Acoustic shadowing, fetal position, and variation in the imaging plane may further affect the visibility of small anatomical targets. Although the controlled perturbation experiments examine selected changes in geometry and visibility, they do not establish performance across these clinical sources of variation. Likewise, the separately trained OCD and AU-OPG models demonstrate applicability to additional imaging modalities rather than direct transfer across ultrasound populations. Future evaluation should include external fetal ultrasound cohorts spanning multiple scanners and gestational ages, with results stratified by image quality and normal versus abnormal cardiac anatomy. This paper addresses static localization. Reader-assistance studies and downstream biometric tasks could clarify its clinical utility [9,37].

6. Conclusions

We presented GRIT-OBB, an oriented method that uses supervised proposal geometry as a shared reference for feature extraction and anatomical localization. The method combines a supervised geometry learner, rotated position-sensitive alignment, a geometry-guided relation residual, and direct overlap supervision at two localization stages. On the FOCUS dataset, GRIT-OBB achieved AP50, AP75, and mAP50:95 of 47.74%, 8.53%, and 16.81%, respectively, corresponding to improvements of 4.74%, 1.8%, and 2.01% over the strongest baseline values. Results on the OCD microscopy dataset and the AU-OPG panoramic dental radiography dataset further suggest that the design generalizes beyond fetal cardiac ultrasound to other medical imaging modalities.

GRIT-OBB also achieved the highest mean mAP50:95 across the 12 geometric and visibility perturbation conditions on each of the datasets: FOCUS, OCD, and AU-OPG. Component ablations support the roles of corrected sampling geometry, regional context, and overlap supervision in improving detection and localization. These findings suggest that correcting proposal geometry before feature extraction and directly supervising both corrected proposals and final predictions provide an effective way to connect target-aligned representations with oriented anatomical localization.

Our code is available at: https://github.com/bill1034/GRIT-OBB (accessed on 28 October 2026).

Author Contributions: Conceptualization, J.Y.; methodology, J.Y.; software, J.Y., W.H., A.C., Q.W., and Z.Y.; validation, J.Y., W.H., A.C., Q.W., and Z.Y.; investigation, W.H., A.C., Q.W., Z.Y., and N.X.; resources, C.F. and F.F.; data curation, W.H., A.C., Q.W., Z.Y., C.F., and F.F.; writing—original draft preparation, J.Y.; writing—review and editing, J.Y., W.H., A.C., Q.W., Z.Y., N.X., and N.H.-C.; supervision, N.H.-C.; project administration, N.H.-C.; funding acquisition, N.H.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Research Development Fund Grant No. RDF-25-01-076.

Institutional Review Board Statement: This study was a secondary analysis of de-identified, publicly available datasets. The study was conducted in accordance with the Declaration of Helsinki. Ethical approval for research involving secondary data was obtained under approval no. ER-LRR-11000211520260102175518.

Informed Consent Statement: Not applicable to the present secondary analysis of de-identified, publicly available data. However, informed consent was obtained from all participants in the original studies as described in the original publications.

Data Availability Statement: The original datasets are available from the sources cited in this article. The additional annotations generated in this study may be available for research purposes upon request to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hernandez-Cruz, N.; Patey, O.; Teng, C.; Papageorghiou, A.T.; Noble, J.A. A comprehensive scoping review on machine learning-based fetal echocardiography analysis. *Comput. Biol. Med.* **2025**, *186*, 109666. https://doi.org/10.1016/j.compbiomed.2025.109666.
2. D'Alberti, E.; Patey, O.; Smith, C.; Šalović, B.; Hernandez-Cruz, N.; Noble, J.A.; Papageorghiou, A.T. Artificial intelligence-enabled prenatal ultrasound for the detection of fetal cardiac abnormalities: A systematic review and meta-analysis. *eClinicalMedicine* **2025**, *84*, 103250. https://doi.org/10.1016/j.eclinm.2025.103250.
3. Mao, Z.; Wang, J.; Zhang, J.; Ohgi, J.; Zheng, Y.; Peng, Y.; Zhao, L.; Su, Q.; Huang, W.; Xu, B. Fine-tuned multimodal large language model for autonomous state cognition system of shape-recognition 6-bar tensegrity integrated with flexible sensors. *Microsyst. Nanoeng.* **2026**, *12*, 228. https://doi.org/10.1038/s41378-026-01322-0.
4. Hernandez-Cruz, N.; Patey, O.; Salovic, B.; Papageorghiou, A.; Noble, J.A. Machine learning-based method for the detection of dextrocardia in ultrasound video clips. *Comput. Methods Programs Biomed.* **2025**, *271*, 109023. https://doi.org/10.1016/j.cmpb.2025.109023.
5. Hernandez-Cruz, N.; Patey, O.; Salovic, B.; Mishra, D.; Sarker, M.M.K.; Papageorghiou, A.; Noble, J.A. Detection of fetal congenital heart defects on three-vessel view ultrasound videos. *WFUMB Ultrasound Open* **2024**, *2*, 100075. https://doi.org/10.1016/j.wfumbo.2024.100075.
6. Hernandez-Cruz, N.; Patey, O.; Adu-Bredu, T.; D'Alberti, E.; Noble, J.; Papageorghiou, A. OP02.04: Automated segmentation of fetal heart three-vessel view ultrasound video clips to facilitate prenatal assessment of congenital heart defects. *Ultrasound Obstet. Gynecol.* **2024**, *64*, 62–63. https://doi.org/10.1002/uog.27886.
7. Hernandez-Cruz, N.; Mishra, D.; Patey, O.; Sarker, M.; Craik, R.; Wilden, E.; Noble, J.A.; Papageorghiou, A.T. OP06.07: Machine learning-based detection of fetal anatomical orientation in second trimester ultrasound images. *Ultrasound Obstet. Gynecol.* **2023**, *62*, 62–62. https://doi.org/10.1002/uog.26499.
8. Mishra, D.; Saha, P.; Zhao, H.; Hernandez-Cruz, N.; Patey, O.; Papageorghiou, A.T.; Noble, J.A. TIER-LOC: Visual Query-based Video Clip Localization in fetal ultrasound videos with a multi-tier transformer. *Med. Image Anal.* **2025**, *103*, 103611. https://doi.org/10.1016/j.media.2025.103611.
9. Huang, C.; Zhang, L.; Xie, B.; Jiang, Y.; Zhu, Y.; Liu, X.; Lei, T.; He, M.; Yan, Y.; Wang, N.; et al. Automated interpretation of fetal cardiac function evaluation from the echocardiogram. *npj Digit. Med.* **2026**, *9*, 334. https://doi.org/10.1038/s41746-026-02381-3.
10. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858. https://doi.org/10.1109/CVPR.2019.00296.
11. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 3163–3171. https://doi.org/10.1609/aaai.v35i4.16426.
12. Han, J.; Ding, J.; Li, J.; Xia, G.S. Align Deep Features for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. 5602511. https://doi.org/10.1109/TGRS.2021.3062048.
13. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3588–3597.
14. Zhao, J.; Ding, Z.; Zhou, Y.; Zhu, H.; Du, W.L.; Yao, R.; El Saddik, A. OrientedFormer: An End-to-End Transformer-Based Oriented Object Detector in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. https://doi.org/10.1109/TGRS.2024.3456240.
15. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162. https://doi.org/10.1109/CVPR.2018.00644.
16. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
17. Arnaout, R.; Curran, L.; Zhao, Y.; Levine, J.C.; Chinn, E.; Moon-Grady, A.J. An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease. *Nat. Med.* **2021**, *27*, 882–891. https://doi.org/10.1038/s41591-021-01342-5.
18. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. https://doi.org/10.1109/ICCV.2017.324.
19. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
20. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3520–3529.

21. Zhao, Z.; Xue, Q.; He, Y.; Bai, Y.; Wei, X.; Gong, Y. Projecting Points to Axes: Oriented Object Detection via Point-Axis Representation. In *Lecture Notes in Computer Science*: Springer: Cham, Switzerland, 2025; Volume 15086, pp. 161–179. https://doi.org/10.1007/978-3-031-73390-1_10.
22. Ding, Z.; Zhou, Y.; Zhao, J.; Du, W.L.; Li, X.; Yao, R.; El Saddik, A. Real-Time Oriented Object Detection Transformer in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 1–14. 5613014. https://doi.org/10.1109/TGRS.2026.3671683.
23. Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; Tian, Q. Rethinking Rotated Object Detection with Gaussian Wasserstein Distance Loss. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; Volume 139, pp. 11830–11841.
24. Marques, J.H.L.; Murrugarra-Llerena, J.; Jung, C.R. GauCho: Gaussian Distributions with Cholesky Decomposition for Oriented Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 3593–3602. https://doi.org/10.1109/CVPR52734.2025.00340.
25. Yuan, J.; Huang, W.; Huang, Y.; Ye, Z.; Shi, Y.; Yang, X.; Xin, N.; Hasan, M.M. DCS-Mamba: Linear-Complexity Spatiotemporal Modeling with Semantic Modulation for High-Resolution Remote Sensing. In *Proceedings of the Advanced Intelligent Computing Technology and Applications*; Huang, D.S., Zhang, C., Chen, W., Li, B., Zhang, Q., Bao, W., Pan, Y., Premaratne, P., Eds.; Springer: Singapore, 2027; Volume 16644, pp. 460–471. https://doi.org/10.1007/978-981-92-3384-7_39.
26. Yuan, J.; Xin, N.; Huang, W.; Wang, Q.; Zhao, L.; Hasan, M.M. KG-STMI: A Knowledge-Driven MultiModal Framework with Triple-Stage Knowledge Injection for Air Quality Forecasting in Environmental Sensing Networks. In *Proceedings of the Knowledge Science, Engineering and Management*; Niu, J., Qiu, M., Cao, C., Eds.; Springer: Singapore, 2027; Volume 16632, pp. 84–99. https://doi.org/10.1007/978-981-92-2852-2_7.
27. Wu, S.; Zhang, H.; Ye, T.; Xie, H.; Zeng, P.; Sun, Q.; Wang, P.; Huang, B.; Du, L.; Wu, G. FOCUS: Four-chamber ultrasound image dataset for fetal cardiac biometric measurement. *Zenodo* **2025**. https://doi.org/10.5281/zenodo.14597550.
28. Kirsten, L.; Angonezi, A.; Marques, J.; Oliveira, F.; Faccioni, J.; Cassel, C.; de Sousa, D.; Vedovatto, S.; Lenz, G.; Jung, C. Oriented Cell Dataset: A Dataset and Benchmark for Oriented Cell Detection and Applications. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 26 February–6 March 2025; pp. 3996–4005.
29. Alabd-Aljabar, A.; Irshaid, Y.; Elsayed, K.; Shaman, M.; Arif, O.; Luke, A.M. Guiding Neural Networks Toward Better Decisions: A Dental Diagnosis and Treatment Planning Framework and Dataset. *IEEE Access* **2026**, *14*, 90109–90121. https://doi.org/10.1109/ACCESS.2026.3703711.
30. Yousef (YSFF). AU-OPG: Panoramic Dental Radiographs With Oriented Tooth-Level Annotations. Hugging Face Dataset. 2025. Available online: https://huggingface.co/datasets/YSFF/AU-OPG (accessed on 27 September 2026).
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
32. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. https://doi.org/10.1109/CVPR.2017.106.
33. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.
34. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
35. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
36. Oksuz, K.; Cam, B.C.; Akbas, E.; Kalkan, S. Localization Recall Precision (LRP): A New Performance Metric for Object Detection. In *Proceedings of the Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; Volume 11211, pp. 521–537. https://doi.org/10.1007/978-3-030-01234-2_31.
37. Huang, W.; Yuan, J.; Wang, Q.; Ye, Z.; Fazi, C.; Fontanella, F.; Hernandez-Cruz, N. Grounding Before Reporting: Local-to-Global Decoupled Learning for Medical Vision-Language Models. In *Advanced Intelligent Computing Technology and Applications*; Springer Nature Singapore: Singapore, 2027; Volume 3035, pp. 291–302. https://doi.org/10.1007/978-981-92-3538-4_25.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.