# DETR-EHPose: An Edge-Enhanced and Heatmap-Guided Framework for DDH Ultrasound Landmark Detection

Zhuofan Wan[1,*], Simiao Tao[1,*], Dandan Zhang[2], Yuxin Wang[1,*]
Qing Zhang[1], Baoying Ye[2], Jiangtao Wang[1], and Hailin Pan[1,†]
[1]Software Engineering Institute, East China Normal University, Shanghai, China
[2]International Peace Maternity and Child Health Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
Emails: {51265902014, 51265902158, 51285902161}@stu.ecnu.edu.cn,
{qzhang, jtwang, hlpan}@sei.ecnu.edu.cn

***Abstract*—Early intervention is critical for ultrasound screening of developmental dysplasia of the hip (DDH), yet speckle noise, weak textures, and blurred boundaries often hinder reliable detection and accurate localization of anatomical landmarks. Although structures such as the cartilage–bone interface and the femoral head contour provide strong edge cues, DETR-style detectors typically emphasize region-level semantics and make limited use of explicit boundary and structural priors. Moreover, Transformer decoders usually rely on iterative refinement, which slows convergence and can degrade performance on small or crowded targets.We propose a structure-aware DETR framework tailored for DDH ultrasound that jointly performs object detection and six keypoint localization. An Edge-aware Feature Propagation (EFP) module extracts multi-scale edge representations from shallow features and strengthens boundary sensitivity via convolutional edge fusion. In addition, we introduce a Heatmap-guided Keypoint Prior (HKP) branch and a Heatmap-guided Query Patch Embedding (HQPE). A coarse keypoint heatmap head on P3 converts sparse keypoint supervision into six-channel Gaussian heatmap regression, providing dense spatial constraints and improved robustness; the features are further enhanced through heatmap-aligned referencing. Experiments on our in-house DDH ultrasound dataset (1,255 training images and 300 test images) demonstrate consistent gains over DETR, DeiM, and strong YOLO baselines in terms of mAP@0.75 and PCK, with notably more stable localization under noise and boundary ambiguity.**

***Index Terms*—Feature fusion, Keypoint localization, RT-DETR, Infant hip ultrasound, Heatmap regression**

## I. Introduction

Developmental dysplasia of the hip (DDH) is a disorder in infants, characterized by acetabular dysplasia, joint instability, and potential dislocation, with an incidence of 1.5–20 per 1,000 neonates [1]. Early screening is critical, as delayed treatment can lead to gait abnormalities, functional impairment, and osteoarthritis. DDH assessment combines physical exams with imaging, with ultrasound preferred due to its accessibility, lack of ionizing radiation, and ability to visualize non-ossified cartilage [2]. The Graf method, the most widely used ultrasound classification for DDH, utilizes $\alpha$ and $\beta$ angle measurements based on anatomical landmarks to assess hip morphology and progression [3], [4].An illustration of these anatomical landmarks and the angle measurements is shown in Fig 1. However, this process is highly operator-dependent, with probe orientation, image quality, and speckle noise affecting accuracy and leading to inter-observer variability [5].

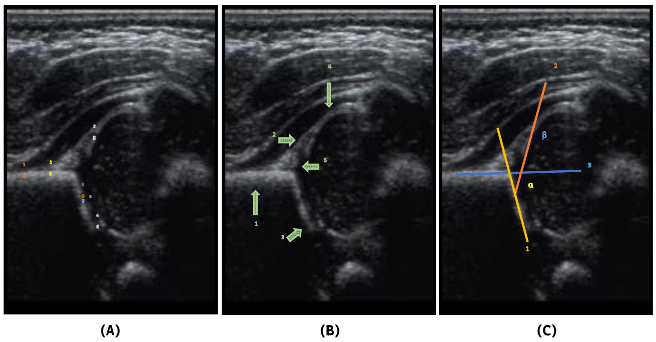


Fig. 1. Key anatomical landmarks on standard coronal hip ultrasound and measurement of Graf classification angles. (A) Schematic illustration of six key landmarks on the standard coronal plane. (B) Structures indicated by arrows: 1, ilium (hyperechoic lateral iliac cortex/iliac baseline); 2, cartilaginous acetabular roof; 3, bony rim/turning point; 4, hip joint capsule; 5, femoral head. (C) Illustration of $\alpha$ and $\beta$ angle measurements for the Graf classification.

Recent advances in deep learning have improved DDH ultrasound assessment: object detection localizes the hip joint and measurement regions, while keypoint localization automates $\alpha/\beta$ angle computation [6], [7]. However, DDH ultrasound images are often affected by speckle noise, weak textures, and blurred boundaries, leading to drift in bounding-box regression and keypoint localization, with a trade-off between accuracy and stability [8], [9]. DETR-style detectors perform end-to-end detection through set prediction and Transformer-based global modeling [10]–[12], excelling in object detection, with

*Equal contribution. †Corresponding author.

TABLE I
DEVELOPMENTAL DYSPLASIA OF THE HIP (DDH) CLASSIFICATION

| Type | $\alpha$ | $\beta$ | Acetabulum Dev. | Clinical Note |
|---|---|---|---|---|
| Type I | $\alpha \geq 60^\circ$ | $\beta < 77^\circ$ | Acetabulum well-developed, femoral head coverage complete | Normal hip joint |
| Type Ia | $\alpha \geq 60^\circ$ | $\beta < 55^\circ$ | Slightly ossified acetabular edge, ossific nucleus triangular | Optimal developmental status |
| Type Ib | $\alpha \geq 60^\circ$ | $55^\circ \leq \beta$ $\beta < 77^\circ$ | Mild ossification delay, may progress to Type II | Need follow-up |
| Type II | $43^\circ \leq \alpha$ $\alpha \leq 59^\circ$ | $\beta < 77^\circ$ | Delayed ossification, increased cartilage coverage | Follow-up, some spontaneous improvement |
| Type IIa | $50^\circ \leq \alpha$ $\alpha \leq 59^\circ$ | $\beta < 77^\circ$ | For infants $<$ 12 months, some may resolve spontaneously | Dynamic observation |
| Type IIb | $50^\circ \leq \alpha$ $\alpha \leq 59^\circ$ | $\beta < 55^\circ$ | For infants $>$ 12 months, ossification delay | Requires treatment (e.g. Pavlik harness) |
| Type IIc | $43^\circ \leq \alpha$ $\alpha \leq 49^\circ$ | $\beta < 77^\circ$ | Poor acetabular development, unstable femoral head | Critical stage, urgent intervention |
| Type IId | $43^\circ \leq \alpha$ $\alpha \leq 49^\circ$ | $\beta > 77^\circ$ | Femoral head "dislocated" and subluxated | Immediate treatment required |
| Type III | $\alpha < 43^\circ$ | $\beta > 77^\circ$ | Femoral head subluxated, ossific nucleus deformed | Requires closed/open reduction |
| Type IIIa | $\alpha < 43^\circ$ | $\beta > 77^\circ$ | Cartilaginous roof hypoechoic (non-fibrotic) | Relatively favorable prognosis |
| Type IIIb | $\alpha < 43^\circ$ | $\beta > 77^\circ$ | Cartilaginous roof hyperechoic (fibrotic) | Poor prognosis |
| Type IV | $\alpha < 43^\circ$ | $\beta > 77^\circ$ | Femoral head completely dislocated, acetabular cavity empty | Requires surgical correction |

RT-DETR [13] improving real-time efficiency. However, their application to ultrasound images is challenging due to their focus on regional semantics and global matching, making them less sensitive to fine-grained edges, while keypoint estimation requires explicit geometric layout constraints not provided by semantic features [14], [15].

This paper proposes a structure-aware DETR framework for DDH ultrasound, incorporating edge enhancement and structural priors to improve robustness and consistency under noisy and blurred conditions. Multi-scale edge representations are extracted from shallow features and integrated into backbone features using edge-preserving downsampling, enhancing the response to key boundaries. Edge and semantic features are fused to improve boundary awareness. A lightweight keypoint heatmap regression head is designed on the $P_3$ feature map to regress Gaussian heatmaps for six keypoints, fused into multi-scale features ($P_3$–$P_5$) [16] to optimize matching. Heatmaps localize candidate centers and keypoints, from which small ROIs are constructed and aligned using ROIAlign [17]. The aggregated patch embedding is fused into the query representation via gated residual fusion, enhancing spatial awareness. The framework automates DDH ultrasound assessment by predicting the measurement region and keypoints, from which $\alpha/\beta$ angles are computed.

The main contributions of this paper are as follows:

- **Edge enhancement and fusion.** Multi-scale edge representations are built using the EFP module and fused with semantic features via ConvEdgeFusion, improving boundary detection and stabilizing bounding-box regression and keypoint localization.
- **Pre-decoder structural prior injection.** A keypoint heatmap head on high-resolution $P_3$ features is supervised by Gaussian heatmaps of six keypoints. Joint optimization of heatmap and detection losses injects a structural prior before the Transformer decoder, enhancing matching and keypoint stability.
- **Heatmap-guided Query Patch Embedding (HQPE).** Center/keypoint heatmaps localize keypoints on $P_3$. Small ROIs are constructed and local patches are extracted via ROIAlign. The aggregated patch embedding $q_{\text{patch}}$ is injected into queries via gated residual fusion, enhancing spatial awareness and region alignment.

## II. RELATED WORK

### A. DDH ultrasound computer-aided assessment

Prior work on DDH ultrasound CAD broadly follows two paradigms. The first focuses on direct diagnosis or classification from images, aiming for end-to-end prediction of DDH types. The second is workflow-oriented and follows the Graf paradigm, typically comprising "standard-plane selection/quality assessment – anatomical localization – angle measurement". Workflow-based systems are often preferred clinically due to interpretability: they explicitly localize anatomical structures and compute $\alpha/\beta$ angles as measurable evidence. The classification of DDH types based on $\alpha$ and $\beta$ angles is summarized in Table I , which provides a comprehensive overview of the different stages and their clinical implications. Recent studies also emphasize robustness to plane variations, acquisition noise, and domain shift as key factors for real-world deployment.

### B. Landmark localization and structural constraints

Heatmap regression is widely used for landmark localization in medical imaging: it converts sparse coordinate supervision into dense Gaussian heatmaps, which improves robustness under noise and ambiguous boundaries. For DDH ultrasound, several works further model landmark dependencies or impose geometric/topological constraints to produce more consistent landmark layouts and reduce jitter. Transformer-based or attention-based designs have also been explored to capture long-range dependencies and mitigate detail loss from downsampling, which is particularly relevant when landmarks are subtle and boundaries are blurred.

### C. Boundary/edge modeling in ultrasound

Edge and boundary cues are critical in ultrasound because anatomical structures may lack strong textures while exhibiting meaningful interfaces (e.g., cartilage–bone). Classic deep edge detectors (e.g., multi-level feature aggregation with deep

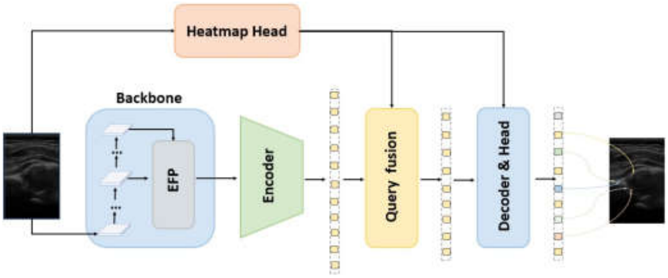


Fig. 2. DETR-EHPose network structure.

supervision) demonstrate the benefits of leveraging shallow-to-deep cues for sharper boundaries. In downstream tasks, injecting edge representations into feature pyramids and using gating/attention to selectively emphasize reliable boundaries can suppress noise-induced pseudo-edges and enhance sensitivity to clinically relevant interfaces.

### *D. DETR-family detectors and priors for localization*

DETR and its variants introduce set prediction with bipartite matching and Transformer encoder–decoder architectures. Deformable attention and real-time adaptations improve convergence and efficiency. Nevertheless, DETR-style detectors often prioritize region-level semantics, leaving room for explicit structural priors in boundary-dominated medical ultrasound. Moreover, query initialization that lacks local geometric evidence can increase the decoder's reliance on iterative refinement. Motivated by these observations, we incorporate multi-scale edge fusion and heatmap-guided landmark priors into a DETR-style framework to improve robustness and stabilize localization in DDH ultrasound.

## III. METHOD

The original RT-DETR does not natively support keypoint detection. To make keypoint supervision compatible with its standard object-detection formulation, we convert each annotated keypoint into a pseudo object by constructing a minimal enclosing bounding box centered at the keypoint (i.e., a small box with a fixed or near-fixed side length). The model is then trained to regress these boxes, and the predicted box centers are used as the corresponding keypoint locations. This conversion enables end-to-end training within the RT-DETR pipeline without modifying its core matching and decoding mechanisms, while preserving the geometric meaning of the original keypoint annotations [10], [13].

### *A. Edge-aware Feature Propagation*

Infant hip ultrasound images often exhibit strong speckle noise, weak textures, and blurred cartilage–bone interfaces, which may cause unstable box regression and landmark localization. To explicitly enhance boundary sensitivity while suppressing noise-induced pseudo-edges, we construct a multi-scale edge pyramid from shallow features and fuse it with semantic features in an edge-aware manner, followed by a lightweight spatial gating module for selective enhancement [18].

Given a shallow feature map $X$, we first extract an explicit edge response using a Sobel operator [19]:

$$E_0 = \mathrm{Sobel}(X). \tag{1}$$

To align with the detection pyramid ($P_3$–$P_5$), we propagate the edge features with edge-preserving downsampling (MaxPool) to obtain multi-scale edge maps [20]:

$$E_{l+1} = \mathrm{MaxPool}(E_l), \quad l \in \{0, 1, \ldots\}. \tag{2}$$

We take $\{E_3, E_4, E_5\}$ and apply $1 \times 1$ convolutions for channel alignment, denoted as $\tilde{E}_i$.

For each pyramid level $i \in \{3, 4, 5\}$ with semantic feature $F_i$, we fuse semantic and edge cues via a lightweight bottleneck (concatenation + $1 \times 1$ conv):

$$Z_i = \phi\Big(\mathrm{Conv}_{1\times1}\big([F_i, \tilde{E}_i]\big)\Big), \tag{3}$$

where $\phi(\cdot)$ denotes a nonlinearity (e.g., ReLU/SiLU).

To avoid indiscriminately amplifying noisy edges, we introduce a spatial gating map computed from channel-wise average/max descriptors [21]:

$$M = \sigma\left(\mathrm{Conv}\left([Z_{\mathrm{avg}}, Z_{\mathrm{max}}]\right)\right) \tag{4}$$

where $\sigma(\cdot)$ is the Sigmoid function and $k$ is typically set to 7.

Finally, we apply gated enhancement and local refinement to produce boundary-aware multi-scale outputs for the RT-DETR decoder:

$$Y_i = \mathrm{Conv}_{1\times1}\Big(\phi\big(\mathrm{Conv}_{3\times3}(M_i \odot Z_i)\big)\Big). \tag{5}$$

The refined features $\{Y_3, Y_4, Y_5\}$ are then fed into the subsequent RT-DETR decoder for detection and keypoint prediction, improving localization stability under noisy and blurred-boundary conditions.

### *B. Heatmap-guided Keypoint Prior Branch*

We introduce a lightweight keypoint heatmap branch into the RT-DETR decoding framework, extending the detection task from "bbox+cls" to "bbox+cls + keypoint heatmap" joint learning. The overall design adheres to the principle of "maintaining real-time performance/low overhead": the new branch runs in parallel to the original Transformer decoding path, without altering the number of queries, attention computation graph, or denoising mechanism, thereby avoiding the introduction of additional sequence length and attention cost.

To supervise the keypoint heatmap branch in the decoder, during data preprocessing we generate a six-channel ground-truth (GT) heatmap for *each* image in the mini-batch, whose spatial resolution is aligned with the $P_3$ feature map [22]. Specifically, the input image is mapped to the heatmap grid using a fixed downsampling stride of $stride = 8$ (i.e., the heatmap size is approximately $1/8$ of the original image), and a zero tensor of shape $[B, 6, H/8, W/8]$ is initialized as the supervision target. For each image, the corresponding landmark

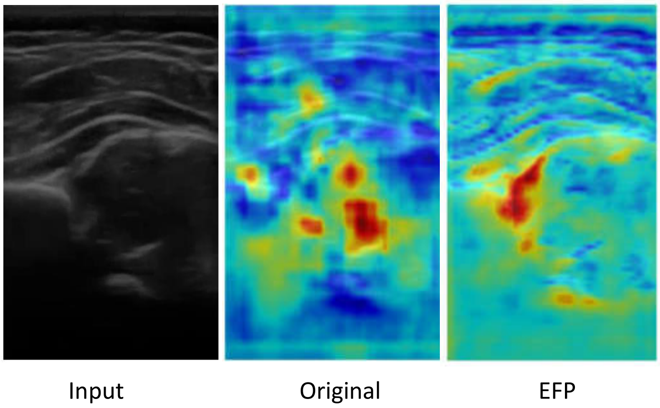


Fig. 3. The first one is the input image, the second is the Layer15 image from the original RT-DETR version, and the third is the Layer16 image under the EFP innovation.

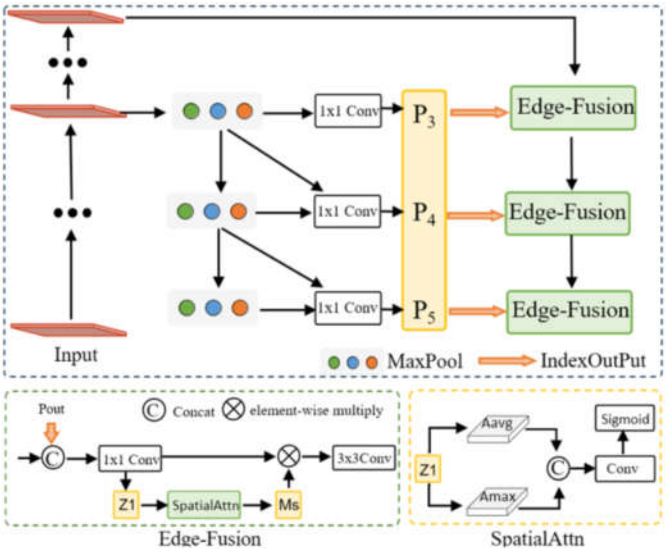


Fig. 4. Edge-aware Feature Propagation.

annotations are projected onto the heatmap coordinate system, and a 2D Gaussian response is rendered at each keypoint location on its respective channel to obtain dense heatmap supervision.

For each image in the batch, we first retrieve the corresponding annotations according to `batch_idx`, and then select the targets belonging to the landmark category (in our implementation, `cls==1`). Here, `cls==1` indicates only that a target is a keypoint instance; the semantic identity of each landmark is encoded by a predefined landmark index $k \in \{1, \ldots, 6\}$, which specifies the heatmap channel. Therefore, we generate a multi-channel heatmap and construct supervision only for these keypoint targets on their corresponding channels.

For each keypoint target, we use the center of its pseudo bounding box as the keypoint location. The center is provided in normalized coordinates (range $[0, 1]$) and is mapped to the heatmap coordinate system to obtain a two-dimensional position $(u, v)$. A 2D Gaussian response (with $\sigma = 2.0$ in heatmap pixels) is then rendered on the heatmap centered at $(u, v)$, transforming the supervision from a single point into a local peak distribution. When multiple keypoints fall at nearby locations on the same channel, we apply a per-pixel "max" fusion strategy:

$$H^{gt}(x, y) \leftarrow \max\Big(H^{gt}(x, y),\ \exp\Big(-\frac{\|p - p_c\|^2}{2\sigma^2}\Big)\Big), \quad (6)$$

where $p = (x, y)$ denotes a pixel location on the heatmap and $p_c = (u, v)$ denotes the Gaussian center.

Given the multi-scale features $\{x_i\}_{i=1}^{L}$ from the backbone, we follow RT-DETR's projection layer to map each layer to a unified hidden dimension $d$. Considering the dependency of keypoint localization on spatial details, we select the highest resolution layer P3 (stride≈8) as the input to the heatmap branch:

$$f_3 = \mathrm{Proj}(x_3) \in \mathbb{R}^{B \times d \times H_3 \times W_3}. \quad (7)$$

Concurrently, a lightweight convolutional head is constructed on $f_3$ to output $K$ channels of heatmap logits, which are then passed through a sigmoid function to obtain the probability heatmap. In practice, the Head consists of two layers of $3 \times 3$ convolution + ReLU and one layer of $1 \times 1$ convolution, featuring a small number of parameters and manageable computational overhead. Additionally, bias initialization is applied to the sparse heatmap: Since keypoint heatmaps are typically highly sparse, we apply a negative bias initialization (e.g., -2.19) to the last convolutional layer's bias, such that the initial $\sigma(b) \approx 0.1$, to suppress background responses and stabilize early training.

A joint loss function is also designed. The overall training objective is the weighted sum of the detection loss and the heatmap loss:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_H \mathcal{L}_{hm}. \quad (8)$$

Here, $\mathcal{L}_{det}$ follows the original RT-DETR loss (including classification/regression and the matching mechanism, while maintaining the computation and aggregation of the denoising branch loss). $\lambda_{hm}$ is the heatmap loss weight (configured via `heatmap_loss_weight` in the implementation) [23].

We provide three optional forms in the implementation (focal/mse/l1), uniformly applying sigmoid to the logits first [24]:

$$L_{hm} = \ell(\sigma(\hat{H}), H^{gt}). \quad (9)$$

When the spatial dimensions of the predicted heatmap $\hat{H}$ and the GT heatmap $H^{gt}$ are inconsistent, bilinear interpolation is applied to $\hat{H}$ to align it to $(H_{hm}, W_{hm})$ before loss calculation, ensuring consistent per-pixel supervision:

$$\hat{H} \leftarrow \mathrm{Interp}(\hat{H}, (H_{hm}, W_{hm})). \quad (10)$$

Furthermore, since simple token/patch embeddings may lose spatial structural information, we enhance regression using

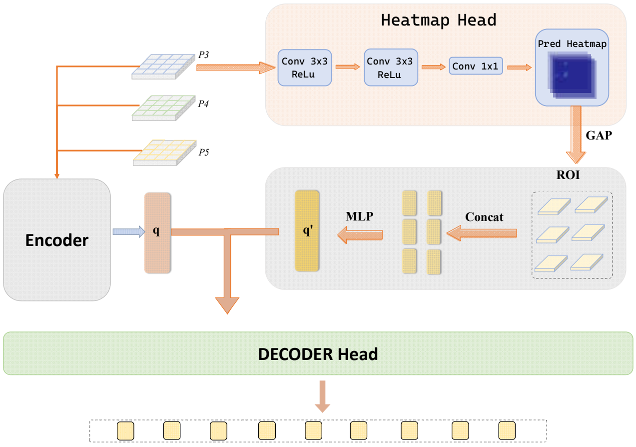


Fig. 5. This figure illustrates the generation process of the heatmap head and the approach for generating bounding boxes under the guidance of heatmap-directed local structural features.

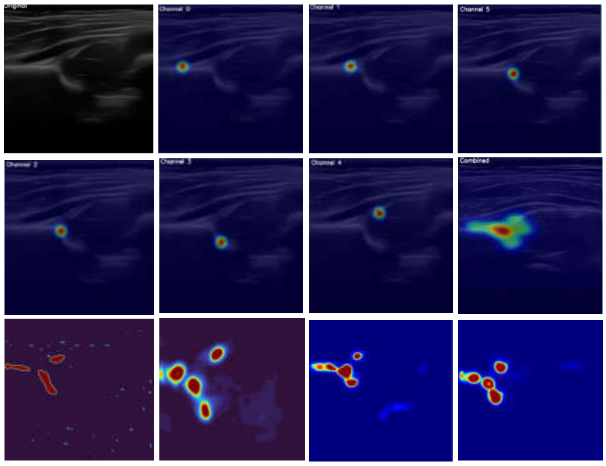

Fig. 6. The images in the first row and second row are the input image and the Gaussian kernel heatmap generated from the six keypoints selected in each batch, respectively. The images in the third row represent the fused predicted heatmaps of the six keypoints generated by the heatmap head at the 50th, 100th, 150th, and 200th epochs, respectively

"structure representation (heatmap) guided local patch embedding".

### C. Heatmap-guided Query Patch Embedding (HQPE)

Existing Transformer-based object detection methods (such as DETR and its variants) typically represent queries as abstract vectors in a global feature space. In RT-DETR, queries are mainly initialized by Top-K selection of encoder features along with anchor priors, where their representation relies more on global semantic information while inadequately modeling the local spatial structure and geometric details around candidate objects.

However, in object detection tasks, accurate bounding box regression highly depends on local texture, edge, and geometric structure information. Relying solely on global feature representations for queries often requires the decoder to iteratively refine localization errors across multiple layers, leading to slower convergence and limited performance in small object or dense scenes.

To address this limitation, we propose a Heatmap-Guided Query Patch Embedding (HQPE) mechanism that enhances query representation by incorporating explicit local structure information guided by keypoint heatmaps.

Specifically, we generate target center heatmaps on the high-resolution feature map P3 using the steps described above, and select Top-K peaks as candidate centers via local non-maximum suppression [25]. Subsequently, leveraging the six keypoint heatmaps, we determine the corresponding locations for each of the six keypoint categories around each candidate center (e.g., by selecting the peak response of each class's heatmap within a local neighborhood of the candidate center). Instead of directly forming a single ROI box from these six points, we treat each keypoint as a small ROI region: we construct a fixed-scale small ROI centered at each keypoint, and crop fixed-size local feature patches (7×7) from P3 via ROIAlign [17]. Since P3 possesses higher spatial resolution, these keypoint-aligned local patches can more fully retain critical information such as object boundaries, texture, and local geometric structures, providing direct evidence for subsequent boundary refinement.

As the decoder queries require one-dimensional vector representations, we first apply global average pooling to each 7×7 two-dimensional patch, converting the six 2D patches into six one-dimensional vectors. These six vectors are then concatenated and passed through a lightweight MLP for dimension reduction and fusion, resulting in a unified local patch embedding denoted as $q_{patch}$. This process effectively aggregates local structural information from the six keypoint locations while ensuring computational efficiency.

During the query fusion stage, we inject the local patch embedding into the original query representation via a gated residual mechanism:

$$q' = q + \alpha \cdot q_{\text{patch}}, \tag{11}$$

where $\alpha$ is an adaptively learned gating coefficient used to dynamically balance the contributions of global semantic features and heatmap-guided local structural features. Through HQPE, queries can obtain local structural supplements aligned with candidate regions before entering the decoder, thereby reducing the decoder's dependence on multi-layer iterative refinement, accelerating convergence, and improving bounding box regression accuracy, especially in small object and crowded scenarios.

## IV. Experiments

### A. Experimental Setup

*1) Dataset and annotation protocol:* We construct an in-house hip ultrasound dataset for developmental dysplasia of the hip (DDH) landmark detection and measurement-region localization. The data were collected at Shanghai International

TABLE II
COMPARISON WITH RECENT REAL-TIME DETECTORS (2024–2026) ON OUR DDH ULTRASOUND TEST SET. RECALL IS COMPUTED AT SCORE=0.3 AND IOU=0.75.

| Method | mAP@0.75 ↑ | Recall@0.75 ↑ | PCK@0.2 ↑ |
|---|---|---|---|
| YOLOv10 (2024) [26] | 79.6 | 88.2 | 83.8 |
| YOLO11 (2024) [27] | 80.8 | 88.6 | 84.6 |
| YOLO12 (2025) [28] | 81.5 | 88.0 | 84.9 |
| RT-DETRv2-R50 (2024) [29] | 84.2 | 83.4 | 81.5 |
| D-FINE-R50 (2024) [30] | 86.0 | 84.6 | 82.7 |
| DEIM + D-FINE (2025) [31] | 86.8 | 85.2 | 83.4 |
| RT-DETRv4-R50 (2025) [32] | 85.5 | 84.1 | 82.3 |
| Ours (EFP + HKP + HQPE) | 88.55 | 89.30 | 89.65 |

TABLE III
ABLATION ON DDH ULTRASOUND TEST SET. BASELINE IS RT-DETR-R50. RECALL IS COMPUTED AT SCORE=0.3 AND IOU=0.75.

| Method | EFP | HKP | HQPE | mAP@0.75 | Recall@0.75 | PCK@0.2 | ΔmAP | ΔRecall | ΔPCK |
|---|---|---|---|---|---|---|---|---|---|
| RT-DETR-R18 | | | | 77.8 | 82.3 | 72.44 | | | |
| RT-DETR-R34 | | | | 80.1 | 83.4 | 78.30 | | | |
| RT-DETR-R50 (baseline) | | | | 83.30 | 80.75 | 80.66 | 0.00 | 0.00 | 0.00 |
| + HQPE | | | ✓ | 85.10 | 86.10 | 83.25 | +1.80 | +5.35 | +2.59 |
| + EFP | ✓ | | | 85.20 | 82.05 | 82.10 | +1.90 | +1.30 | +1.44 |
| + HKP | | ✓ | | 84.40 | 82.70 | 84.05 | +1.10 | +1.95 | +3.39 |
| + EFP + HKP | ✓ | ✓ | | 86.10 | 84.55 | 85.55 | +2.80 | +3.80 | +4.89 |
| + HKP + HQPE | | ✓ | ✓ | 86.45 | 88.60 | 86.85 | +3.15 | +7.85 | +6.19 |
| + EFP + HQPE | ✓ | | ✓ | 87.10 | 88.05 | 86.20 | +3.80 | +7.30 | +5.54 |
| Ours (+ EFP + HKP + HQPE) | ✓ | ✓ | ✓ | **88.55** | **89.30** | **89.65** | **+5.25** | **+8.55** | **+8.99** |

Maternal and Child Health Hospital between Sep. 2024 and Nov. 2025 during routine clinical B-mode hip ultrasound examinations. The dataset contains 723 subjects and 1,555 2D ultrasound images (including frames extracted from ultrasound videos).

Following the Graf workflow, each image is annotated with: (i) a measurement region bounding box covering the valid area for $\alpha/\beta$ angle construction; and (ii) six anatomical landmarks used to compute the Graf angles (Fig. 3). For training in an RT-DETR pipeline, each landmark is represented as a pseudo object by a small square bounding box centered at the landmark location, enabling unified bipartite matching and end-to-end learning [33].

All annotations were produced by experienced clinicians and further reviewed by a senior chief expert for consistency and quality control. The dataset is split by subject to avoid patient leakage, with 1,255 images for training and 300 images for testing. All data were de-identified before analysis. The study was approved by the Institutional Review Board of Shanghai International Maternal and Child Health Hospital, and written informed consent was obtained from the participants' legal guardians (IRB No. GKLW-A-2024-040-01).

*2) Preprocessing and augmentation:* We remove screen overlays and non-imaging regions when present, and resize all images to $640 \times 640$. Standard augmentations are applied during training, including random flipping, rotation, and scaling, to improve robustness to probe orientation and acquisition variability.

*3) Compared methods:* We compare our method against representative real-time detectors and DETR-family variants:

- RT-DETR with different backbones: ResNet-18/34/50 (denoted as RT-DETR-R18/R34/R50) [33].
- YOLO series as strong one-stage real-time baselines: YOLOv10/YOLO11/YOLO12 [26]–[28].
- Recent DETR-family real-time variants: RT-DETRv2-R50, D-FINE-R50, DEIM + D-FINE, and RT-DETRv4-R50 [29]–[32].
- RT-DETR + HKP: RT-DETR augmented with our heatmap-guided keypoint prior (HKP) supervision branch.
- RT-DETR + HQPE: RT-DETR augmented with our Heatmap-guided Query Patch Embedding.
- RT-DETR + EFP: RT-DETR augmented with our Edge-aware Feature Propagation.
- Ours (EFP + HKP + HQPE): RT-DETR with EFP, HKP, and HQPE enabled.

All methods are trained and evaluated under the same input resolution, data splits, and evaluation protocol for fairness.

*4) Evaluation metrics:* We evaluate (1) measurement-region detection and (2) landmark localization. For detection, we report mAP@0.75 (AP at IoU threshold 0.75) and Recall@0.75 computed at score=0.3 and IoU=0.75, following the standard AP/AR definition used in COCO-style evaluation [34]. For landmark localization, we compute PCK@0.2 [35], where a predicted landmark is considered correct if the normalized

Euclidean error is below a fraction $\tau$:

$$\mathrm{PCK}@\tau = \frac{1}{NK}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{I}\Big(\|\hat{\mathbf{p}}_{n,k} - \mathbf{p}_{n,k}\|_2 \le \tau \cdot s_n\Big), \quad (12)$$

where $K$=6 landmarks, $\hat{\mathbf{p}}_{n,k}$ and $\mathbf{p}_{n,k}$ denote predicted and ground-truth landmark coordinates, and $s_n$ is a per-image normalization term derived from the measurement-region size (we use the diagonal length of the ground-truth measurement-region box). The predicted landmark coordinate is taken as the center of the predicted pseudo-box.

*5) Implementation details:* All models are trained end-to-end with AdamW [36]. We use a learning rate of $2 \times 10^{-4}$ for the backbone and $2 \times 10^{-3}$ for newly introduced modules (EFP/HKP/HQPE/heatmap head), with gradient clipping set to 0.1. Unless otherwise specified, other hyperparameters (e.g., number of queries, denoising, loss weights) follow the official RT-DETR setting [33]. The HKP heatmap branch operates on the $P_3$ feature (stride $\approx 8$), predicts $K$=6 Gaussian heatmaps, and uses $\sigma$=2.0 for GT heatmap generation. In HQPE, local patches are cropped by ROIAlign from $P_3$ using a $7\times7$ pooling size and fused into queries via gated residual fusion.

### B. Comparison with Baselines

Table II summarizes the quantitative comparison with recent real-time detectors on our test set. Among all listed methods, our full model (EFP + HKP + HQPE) achieves the best overall performance, reaching 88.55 mAP@0.75, 89.30 Recall@0.75, and 89.65 PCK@0.2.

Compared with the RT-DETR-R50 baseline in Table III, our method improves mAP@0.75 from 83.30 to 88.55 (+5.25), Recall@0.75 from 80.75 to 89.30 (+8.55), and PCK@0.2 from 80.66 to 89.65 (+8.99).

One-stage baselines. The YOLO series shows strong recall (e.g., YOLO12 reaches 88.0 Recall@0.75 in Table II) but lower mAP@0.75 than our method under the strict IoU criterion. In particular, our method improves mAP@0.75 by +7.05 (81.5→88.55) and PCK@0.2 by +4.75 (84.9→89.65) over YOLO12, indicating tighter measurement-region localization and more accurate landmark prediction.

DETR-family baselines. Recent DETR-style real-time variants (e.g., D-FINE-R50 and DEIM + D-FINE) achieve stronger mAP than YOLO baselines, yet our approach still yields consistent gains in both detection and keypoint localization, suggesting that boundary-aware propagation and heatmap-aligned query patch injection are beneficial for boundary-ambiguous ultrasound.

### C. Ablation Study

We conduct ablations to isolate the contribution of each proposed component (EFP, HKP, HQPE) and their combinations. All ablations use the same training protocol and evaluation metrics, and results are summarized in Table III.

HQPE. Adding HQPE improves both detection and keypoint localization (RT-DETR-R50→RT-DETR-R50+HQPE), increasing mAP@0.75 from 83.30 to 85.10 and Recall@0.75 from 80.75 to 86.10, while improving PCK@0.2 from 80.66 to 83.25. This indicates that heatmap-guided local patch evidence strengthens query representations and improves matching and regression quality.

HKP. Adding HKP alone improves PCK@0.2 to 84.05, suggesting that dense heatmap supervision provides a stable structural signal for landmark learning, while also yielding modest gains in detection.

EFP. EFP improves mAP@0.75 to 85.20, indicating that explicit boundary enhancement benefits tighter region regression under strict IoU, while also providing moderate gains in Recall@0.75 and PCK@0.2.

Complementarity. Combining components yields larger improvements than any single module. For example, HKP + HQPE substantially boosts recall (88.60) and PCK (86.85), while EFP + HQPE improves mAP@0.75 to 87.10. Enabling all three components achieves the best overall result (88.55 mAP@0.75 / 89.30 Recall@0.75 / 89.65 PCK@0.2), demonstrating that edge-enhanced features (EFP), dense structural supervision (HKP), and heatmap-aligned query patches (HQPE) are complementary.

### D. Qualitative visualization of the heatmap prior

To inspect the behavior of the heatmap-guided design, Fig. 6 visualizes the GT Gaussian heatmaps and the predicted heatmaps at different epochs. The predicted responses gradually evolve from diffuse activations to compact peaks aligned with clinically relevant structures, indicating that dense heatmap supervision provides a stable structural signal even under speckle noise and boundary ambiguity.

## V. Conclusion

This paper presents a structure-aware RT-DETR framework for infant hip ultrasound in developmental dysplasia of the hip (DDH), targeting robust measurement-region detection and six-keypoint localization under speckle noise and blurred cartilage–bone boundaries. We incorporate three lightweight modules—Edge-aware Feature Propagation (EFP), a Heatmap-guided Keypoint Prior (HKP) branch, and Heatmap-guided Query Patch Embedding (HQPE)—to inject boundary cues and structural priors into DETR-style decoding, while preserving the real-time design philosophy. Experiments and ablations on an in-house DDH dataset demonstrate improved detection and landmark localization stability, supporting automatic computation of clinically relevant Graf $\alpha/\beta$ angles. Future work will extend validation to larger multi-center cohorts and investigate uncertainty-aware and domain-adaptive strategies for safer clinical deployment.

## References


[1] J. Krauss, M. Straus Takahashi, and J. Samet, "Developmental dysplasia of the hip: ultrasound evaluation," *Pediatric Radiology*, vol. 55, no. 10, pp. 1978–1991, 2025.

[2] M. A. Taylor and J. G. Maclean, "Anterior hip ultrasound: A useful technique in developmental dysplasia of the hips," *Ultrasound*, vol. 29, no. 3, pp. 179–186, 2021.

[3] R. Graf, "The diagnosis of congenital hip-joint dislocation by the ultrasonic combound treatment," *Archives of Orthopaedic and Traumatic Surgery*, vol. 97, no. 2, pp. 117–133, 1980.

[4] M. Wako, H. Kono, K. Koyama, T. Fujimaki, N. Furuya, H. Haro, and M. Wako, "The anatomical position of graf's standard plane and its relationship with pelvic morphology: a computed tomography-based study," *Cureus*, vol. 14, no. 7, 2022.

[5] A. Roposch, R. Graf, and J. G. Wright, "Determining the reliability of the graf classification for hip dysplasia," *Clinical Orthopaedics and Related Research*, vol. 447, pp. 119–124, 2006.

[6] M. Kinugasa, A. Inui, S. Satsuma, D. Kobayashi, R. Sakata, M. Morishita, I. Komoto, and R. Kuroda, "Diagnosis of developmental dysplasia of the hip by ultrasound imaging using deep learning," *Journal of Pediatric Orthopaedics*, vol. 43, no. 7, pp. e538–e544, 2023.

[7] H. Shimizu, K. Enda, H. Koyano, T. Ogawa, D. Takahashi, S. Tanaka, N. Iwasaki, and T. Shimizu, "Diagnosis on ultrasound images for developmental dysplasia of the hip with a deep learning-based model focusing on signal heterogeneity in the bone region," *Diagnostics*, vol. 15, no. 4, p. 403, 2025.

[8] S.-W. Lee, H.-U. Ye, K.-J. Lee, W.-Y. Jang, J.-H. Lee, S.-M. Hwang, and Y.-R. Heo, "Accuracy of new deep learning model-based segmentation and key-point multi-detection method for ultrasonographic developmental dysplasia of the hip (ddh) screening," *Diagnostics*, vol. 11, no. 7, p. 1174, 2021.

[9] S. Taşolar, M. A. Günen, A. Sığırcı, and H. Taşolar, "A deep learning–based framework for automatic determination of developmental dysplasia of the hip from graf angles," *Journal of Imaging Informatics in Medicine*, pp. 1–15, 2025.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[11] T. Huang, J. Shi, J. Li, J. Wang, J. Du, and J. Shi, "Involution transformer based u-net for landmark detection in ultrasound images for diagnosis of infantile ddh," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 8, pp. 4797–4809, 2024.

[12] T. Huang, J. Shi, G. Jin, J. Li, J. Wang, Q. Wang, J. Du, and J. Shi, "Topological gcn guided improved conformer for detection of hip landmarks from ultrasound images," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[13] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 965–16 974.

[14] J. Huang and T. Li, "Small object detection by detr via information augmentation and adaptive feature fusion," in *Proceedings of 2024 ACM ICMR Workshop on Multimodal Video Retrieval*, 2024, pp. 39–44.

[15] D. Kumar, M. A. Mehta, H. Müller, and A. Holzinger, "Transformer-powered precision: A detr-based approach for robust detection in medical ultrasound with cholelithiasis as a case study," *Computational and Structural Biotechnology Journal*, vol. 28, pp. 454–467, 2025.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[18] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5229–5238.

[19] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," Presented at the Stanford Artificial Intelligence Project (SAIL), 1968. [Online]. Available: https://www.semanticscholar.org/paper/An-Isotropic-3%C3%973-image-gradient-operator-Sobel-Feldman/1ab70add6ba3b85c2ab4f5f6dc1a448e57ebeb30

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. M. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[22] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 466–481.

[23] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[26] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[27] Ultralytics, "Ultralytics yolo11 documentation," https://docs.ultralytics.com/models/yolo11/, 2024, accessed 2026-02-01.

[28] Y. Tian, Q. Ye, and D. Doermann, "Yolov12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[29] W. Lv, Y. Zhao, Q. Chang, K. Huang, G. Wang, and Y. Liu, "Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer," *arXiv preprint arXiv:2407.17140*, 2024.

[30] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-fine: Redefine regression task in detrs as fine-grained distribution refinement," *arXiv preprint arXiv:2410.13842*, 2024.

[31] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "Deim: Detr with improved matching for fast convergence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[32] Z. Liao, Y. Zhao, X. Shan, Y. Yan, C. Liu, L. Lu, X. Ji, and J. Chen, "Rt-detrv4: Painlessly furthering real-time object detection with vision foundation models," *arXiv preprint arXiv:2510.25257*, 2025.

[33] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," *arXiv preprint arXiv:2304.08069*, 2024.

[34] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[35] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3686–3693.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.