Topological GCN Guided Improved Conformer for Detection of Hip Landmarks from Ultrasound Images

Tianxiang Huang, Jing Shi, Ge Jin, Juncheng Li, Jun Wang, *Member, IEEE*, Qian Wang, Jun Du, and Jun Shi, *Member, IEEE*

***Abstract*—The B-mode ultrasound based computer-aided diagnosis (CAD) has shown its effectiveness for diagnosis of Developmental Dysplasia of the Hip (DDH) in infants within 6 months. Hip landmark detection is a feasible way for the CAD of DDH according to the Graf's method. However, existing landmark detection algorithms mainly focus on designing special models to capture the features from hip ultrasound images, but generally ignore the important spatial relations among different landmarks. To this end, a novel weakly supervised learning-based algorithm, the Topological Graph Convolutional Network (TGCN) guided Improved Conformer (TGCN-ICF), is proposed for detecting landmarks from hip ultrasound images. The TGCN-ICF includes two subnetworks: an Improved Conformer (ICF) subnetwork to generate heatmaps and constraint vectors from ultrasound images, and a TGCN subnetwork to additionally explore topological relations among hip landmarks with the guidance of class labels for further refining and improving the detection accuracy. Moreover, a new Mutual Modulation Fusion (MMF) module is developed to fully exchange and fuse the extracted feature information from the convolutional neural network (CNN) and Transformer branches in ICF. Meanwhile, a novel Mutual Supervision Constraint (MSC) strategy is designed to provide a constraint for detection of each hip landmark. The experimental results on two real-world DDH datasets demonstrate that the TGCN-ICF outperforms all the compared algorithms, suggesting its potential applications. The source code is publicly available on https://github.com/Tianxiang-Huang/TGCN-ICF.**

***Index Terms*—Developmental Dysplasia of the Hip (DDH), B-mode Ultrasound Images, Landmark Detection, Topological Graph Convolutional Network.**

I. Introduction

DEVELOPMENTAL Dysplasia of the Hip (DDH) is one of the most common orthopedic disorders in infants, which may result in acetabular dysplasia, hip instability, and hip dislocation [1]. It is crucial to accurately diagnose DDH in the early stagy for the following treatment [2]. In clinical practice, B-mode Ultrasound (BUS) imaging is commonly used for diagnosis of DDH in infants within 6 months, due to the advantages of non-invasive, non-radiation, and real-time imaging [3]. The Graf's method is considered as the gold standard examination for diagnosing DDH by measuring the α and β angles as shown in Fig. 1(a) [4]. However, this method is susceptible to the subjective expertise of sonologists. Therefore, the computer-aided diagnosis (CAD) for DDH has gained its reputation in recent years.

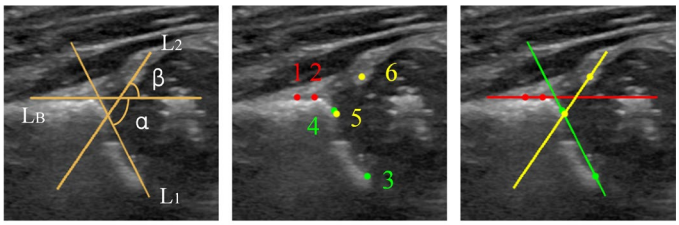

Fig. 1. Illustration of hip BUS images according to Graf's method. (a) Two angles. α is formed by the angle between the base line (L_B) and the bone roof line (L_1), β is created by the intersection of the base line (L_B) and the cartilage roof line (L_2). (b) Six landmarks. Landmarks 1 to 6 are the anatomical critical points in the hip BUS images. (c) Three lines. The red base line (L_B) is formed by landmark 1 and 2, the green bone roof line (L_1) is formed by landmark 3 and 4, and the yellow cartilage roof line (L_2) is formed by landmark 5 and 6.

Recently, deep learning (DL) has garnered significant attention in the field of BUS-based CAD for DDH [5], [6], [7], [8]. Most of these models focus on developing specialized segmentation algorithms to segment the critical anatomical structures for calculating two angles [5], [6], resulting in the issue that the accuracy of angle measurement extremely depends on the performance of segmentation algorithms. In fact, the angle measurement can be simply determined by some critical hip landmarks as shown in Fig. 1(b) and (c). Some pioneering works then have explored the feasibility of

This work was supported in part by the National Natural Science Foundation of China under Grant 62271298, in part by the 111 Project under D20031, and in part by the Fund of the Cyrus Tang Foundation, and the Fund of the Education Development Foundation of Shanghai Jiao Tong University. (*Corresponding author: Jun Shi and Jun Du*).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by IRB of Shanghai Children's Medical Center Affiliated to Shanghai Jiao Tong University School of Medicine under Application No. SCMCIRB-K2023027-1.

T. Huang, G. Jin, J. Li, J. Wang, and J. Shi are with the Key Laboratory of Specialty Fiber Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. (e-mail: junshi@shu.edu.cn).

J. Shi, Q. Wang, and J. Du are with the Imaging Diagnosis Center, Shanghai Children's Medical Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200127, China. (e-mail: dujun@scmc.com.cn).

developing the CAD based on hip landmark detection [7], [8]. However, it is still a challenging task to accurately detect the critical hip landmarks due to the effect of speckle noise in ultrasound images [9].

As shown in Fig. 1(b), the hip landmarks are distributed across multiple regions in BUS images. Thus, it is essential to capture both the local and global information to improve the detection accuracy of DL-based detection models. Since the convolutional neural network (CNN) mainly focuses on extracting local features [10], while the Transformer architecture can well learn global representations [11], the hybrid models by combining CNN and Transformer then indicate their superior performance for the point detection task [12]. As a classical hybrid model, Conformer is specially designed a dual-branch structure to integrate the CNN with the visual Transformer into a unified framework. Moreover, the Conformer model has shown its effectiveness in many computer vision tasks [13]. Therefore, it is feasible to adopt the Conformer as the backbone network in our hip landmark detection task. However, the simple features fusion strategy in Conformer, such as concatenation, cannot fully fuse the local and global features extracted from two branches, which will affect the detection performance to a certain extend.

Moreover, existing heatmap generation-based landmark detection methods generally hypothesize that the landmarks are independent [14], and thus they only generate the corresponding heatmaps by the neighborhood information surrounding with the points. In fact, the hip landmarks in the BUS images are correlated with each other, or they exhibit some special spatial relations based on topological position information. For example, as shown in Fig. 1(b), the red landmarks (landmark 1 and 2) are collinear, which serve as the key points to form the base line (L_B in Fig. 1(a)) according to the Graf's method [4]. These special relations among different landmarks can provide important spatial topology knowledge and constraint to help enhance the detection accuracy. However, existing hip landmark detection algorithms do not pay close attention to this important prior and constraint information, and it is also difficult to model and utilize them.

In recent years, graph convolutional network (GCN) has gain its reputation to learn informative representations for different tasks in medical image analysis [15]. When constructing a graph, each node has its inherent features, and the graph edges can represent the special relations of different nodes [16]. The GCN is one of the most representative neural network-based graph learning methods, which can effectively model rich relational information by using convolution operation on the graph [17]. Thus, it is necessary to further apply the GCN to capture valuable prior spatial relationships among different landmarks for refining and improving detection performance. Since each hip ultrasound image can be annotated with an image-level diagnostic class label, it is then feasible to use the class labels for training graph convolutional network, so as to capture and learn the spatial topological information. Here, the image-level labels are not directly related to the labeled landmarks, and therefore, it is a weakly supervised-based approach for landmark detection.

Besides the existing spatial topological relations based on the Graf's method, each detected point within the hip BUS images can be further constrained and supervised by other landmarks. For example, as illustrated in Fig. 1(b), the landmark 2 consistently appears to the right of the landmark 1, and is simultaneously in the upper-left corner relative to the landmark 3. These specific spatial relationships across different landmarks can provide additional supervisory information to guide the performance of hip landmark detection. However, the previous point detection algorithms generally ignore these meaningful spatial constraint relations. Therefore, it is feasible to provide more supervision for each point, so as to further promote the hip landmark detection.

In this work, we propose a novel weakly supervised learning-based algorithm, namely TGCN-ICF, for hip landmark detection in ultrasound images. The TGCN-ICF consists of two subnetworks: an Improved Conformer (ICF) subnetwork to generate the related heatmaps and constraint vectors, and a Topological GCN (TGCN) subnetwork to further refine landmark detection with the guidance of class labels in the paradigm of weakly supervised learning. Moreover, a Mutual Modulation Fusion (MMF) module is developed to fully exchange and fuse features extracted from the CNN and Transformer branches in ICF. Furthermore, a Mutual Supervision Constraint (MSC) strategy is designed to model the constraint relationships among the hip landmarks to improve detection. The experimental results on two real DDH BUS datasets indicate the effectiveness of the proposed TGCN-ICF.

The main contributions of this work are summarized as follows:

1) A novel weakly supervised TGCN-ICF algorithm is proposed for hip landmark detection from BUS images. Different from the conventional heatmap generation-based detection approaches, the class labels are innovatively applied as the weakly supervised information to guide the learn of the specially designed TGCN subnetwork, which models the valuable topological relations among hip landmarks into a graph based on the inherent properties of landmarks in ultrasound images. Thus, the TGCN can further refine the generated heatmaps for effectively improving detection accuracy.
2) A new MMF module is developed in the ICF subnetwork to fully exchange and fuse the local and global features that extracted from the CNN and Transformer branches, respectively. In particular, the local and global features are adaptively enhanced by learning information from each other in MMF, which then can effectively improve feature representation of the ICF subnetwork for subsequently generating related heatmap of each landmark.
3) A novel MSC strategy is designed to explore the spatial constraints among different hip landmarks. Specifically, the MSC strategy imposes the relative position supervision and constraints on each hip point by the coordinate difference vectors, so as to implicitly correct the detection deviations by modeling the correlations among these landmarks during model training.

This paper is an extension of our work published in MICCAI 2024 [18]. The MSC strategy is further proposed to be integrated into TGCN-ICF, which provides additional constraints for each landmark via the coordinate regression-based approach, thereby enhancing the detection performance. In addition, more details are introduced about the proposed TGCN-ICF, and more experiments are conducted on two real-world DDH datasets.

II. Related Works

A. DL-based Methods for BUS DDH Diagnosis

The DL-based approaches have indicated their effectiveness for BUS-based CAD of DDH. According to the Graf's method, existing CAD models are mainly developed based on measuring the α and β angels, which can be divided into two categories: the segmentation-based model and the landmark detection-based model. The former mainly segments the critical anatomical structures, while the latter directly detect the key points within BUS images.

It is worth noting that most CADs for DDH focus on developing special segmentation algorithms for the anatomical structures to perform the followed angle measurement. For instance, Hu et al. [19] proposed a multi-task learning network by using the Mask R-CNN as the basic framework for the detection and segmentation of four anatomical structures in hip ultrasound images; Liu et al. [5] developed an attention-based segmentation network, named NHBS-Net, to segment seven key structures of the neonatal hip joint. All these works demonstrate the success of segmentation-based approaches.

However, the training of segmentation-based CAD models requires professional and laborious annotation, and generally suffers from the issue of small sample size. On the contrary, the landmark detection-based method is more convenient, since it only needs to annotate several key points in the hip BUS images. Some works then have developed the effective point detection algorithms for diagnosis of DDH. For example, Xu et al. [7] proposed a Dependency Mining ResNet (DM-ResNet) to capture both short-range and long-range dependencies for detecting six hip landmarks from ultrasound images; Huang et al. [8] proposed an IT-UNet network that integrates involution operation into Transformer to capture both spatial-related and long-range information for detecting critical landmarks within BUS images. These previous works indicate the feasibility and effectiveness of hip landmark detection for DDH diagnosis.

Existing hip landmark detection algorithms primarily focus on designing special DL detection models, and do not attach importance to the prior knowledge about the spatial relationships. In this work, we aim to explore the additional prior information, such as the topological relations and spatial constraints among hip landmarks, so as to further enhance the detection performance of hip landmarks.

B. Topological Relations Guided GCN

The topological relations (e.g., adjacency, inclusion, and exclusion) are the reliable and valuable information hidden in the images, which can be served as the additional guidance for DL models, so as to enhance their performance [20]. Some recent works have explored and modeled the topological interactions in images, and suggested the feasibility and effectiveness [21], [22]. Moreover, since the GCN has a strong ability in modeling and representing relational information [23], it is also considerable to integrate the topological information into a graph for learning graph representation. For example, Zhang et al. [24] proposed a new topological graph segmentation model for lung tumor segmentation, which integrated the topological features into the graph convolutional layers for improving the segmentation performance; Wang et al. [25] developed a novel topology-aware Transformer network for 3D hand pose estimation, which captured both the long-range dependencies and local topology connection by deeply integrating Transformer and GCN layers; Shi et al. [26] introduced a novel topology-aware hybrid architecture that employed the Pool GNN module and Swin GNN module to learn both global and local topological representations for complex anatomical structures segmentation in medical images.

It is worth noting that there are also some inherent relations among different landmarks within hip BUS images according to the Graf's method [4]. Therefore, we propose an additional TGCN subnetwork to further learn the valuable topological graph representations with the guidance of class labels.

C. Feature Fusion in Hybrid Model

Recently, the CNN-Transformer based hybrid models have shown their effectiveness in the field of medical image analysis [27], since these models can effectively explore both the local and global information by combining the complementary strengths of these two architectures. However, it is still a challenging task to fully fuse and leverage the advantages of both CNN and Transformer branches. To this end, Zhang et al. [28] designed a BiFusion module that incorporated both self-attention and multi-modal fusion mechanisms to efficiently fuse the multi-level features from both network branches; Dai et al. [29] proposed an attentional feature fusion module and an iterative attentional feature fusion module based on multi-scale channel attention to better fuse features of different network branches; Liu et al. [30] developed a feature super decoder to effectively fuse the multi-level features of both branches, and further designed a multi-scale feature aggregation module to complement the location and spatial information. These works demonstrate the importance of feature fusion in hybrid models.

However, existing fusion strategies are mainly developed by the simple concatenation, redundant convolutions, or complex attention mechanisms. Different from these approaches, we design a MMF module to deeply exchange and fuse the features extracted from CNN and Transformer branches, which can adaptively enhance each branch's information learning by another branch.

III. Method

As shown in Fig. 2, the proposed TGCN-ICF algorithm for hip landmark detection consists of two subnetworks, namely an ICF subnetwork and a TGCN subnetwork. Moreover, a MSC

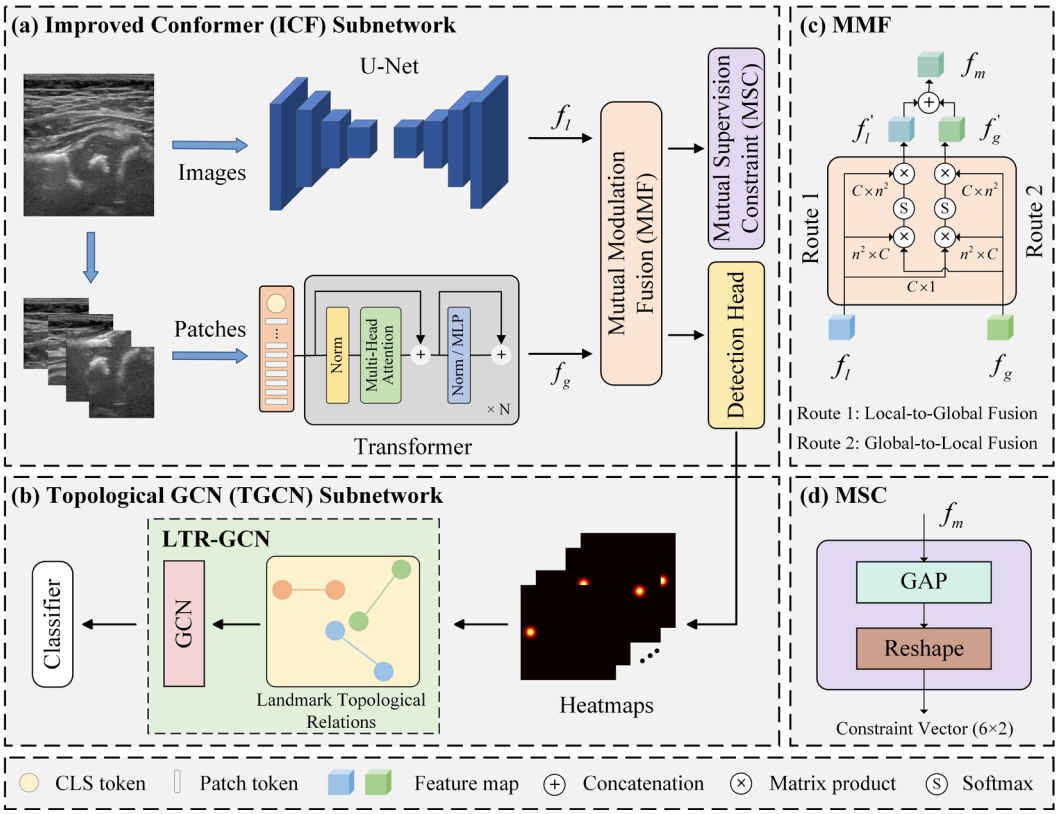

Fig. 2. Overview of the proposed Topological GCN guided Improved Conformer (TGCN-ICF). (a) Improved Conformer (ICF) Subnetwork. (b) Topological GCN (TGCN) Subnetwork. (c) Mutual Modulation Fusion (MMF) module. (d) Mutual Supervision Constraint (MSC) module.

strategy is designed for further improvement of landmark detection. The training pipeline of TGCN-ICF is as follows:

1) The hip BUS images and the corresponding patches are first fed into the ICF subnetwork to generate heatmaps and constraint vectors.
2) The generated heatmaps are then fed into the TGCN subnetwork for further refinement with the guidance of class labels. Moreover, the generated constraint vectors are utilized to constraint each landmark by the MSC strategy.

The details of TGCN-ICF are then introduced in the following section, mainly including the ICF Subnetwork, the TGCN Subnetwork, the MSC Strategy, and the Loss Function.

A. ICF Subnetwork

In the ICF subnetwork, since U-Net is a commonly used encoder-decoder architecture for landmark detection task, we replace the conventional CNN branch in the original Conformer with the U-Net. Thus, the U-Net and Transformer branches can effectively capture both the local and global information in hip BUS images. Meanwhile, inspired by the works in [31], a MMF module is developed to deeply exchange and fuse the features extracted from these two branches. The MMF module can adaptively update and optimize each branch's information by another branch, achieving highly effective fusion of local and global features.

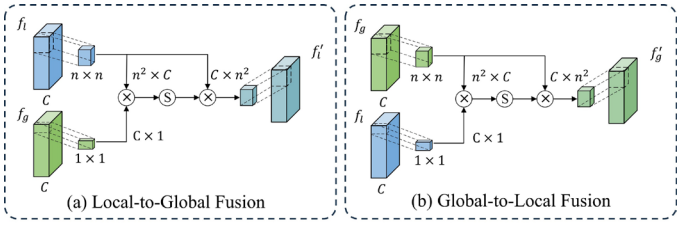

Fig. 3. The detailed illustration of the MMF module. (a) Local-to-Global Fusion route; (b) Global-to-Local Fusion route.

As shown in Fig. 3, define two feature maps $f_l \in \mathbb{R}^{h\times w\times c}$ and $f_g \in \mathbb{R}^{h\times w\times c}$ that are extracted from the U-Net branch and Transformer branch, respectively. To fuse f_l and f_g, we specifically design two synchronous fusion routes: Local-to-Global Fusion and Global-to-Local Fusion.

1) Local-to-Global Fusion

In this Local-to-Global Fusion route, the f_l is updated by f_g in the pixel level. Specifically, a filter $F^{lg}(\cdot)$ is learned to update the local neighbor pixels (denoted as $L^{n^2}_{(i,j)}$) in an $n \times n$ neighborhood with the corresponding center pixel $G_{(i,j)}$ in f_g. The filter weight is defined as follows:

$$\boldsymbol{\omega}^0_{(i,j)} = softmax\left(\sum\nolimits_c (G_{(i,j)} \otimes L^{n^2}_{(i,j)})\right) \quad (1)$$

where $softmax(\cdot)$ represents the normalized exponential function, $\sum_c(\cdot)$ denotes the summation along the channel dimension, and $\otimes$ is the matrix product. Therefore, the updated neighbor pixels can be calculated by:

$$L^{n^2}_{(i,j)}{}' = F^{lg}\left[L^{n^2}_{(i,j)}\right] = \sum\nolimits_{n^2} (L^{n^2}_{(i,j)} \otimes \boldsymbol{\omega}^0_{(i,j)}) \quad (2)$$

where $L^{n^2}_{(i,j)}{}'$ denotes the updated local pixels in the $n \times n$ neighborhood, and $\sum_{n^2}(\cdot)$ represents the summation along the neighborhood spatial dimension. Thus, all pixels in f_l are updated by targeting the counterpart pixels in f_g, and then we

can obtain the fused local-to-global information f_l'.

2) Global-to-Local Fusion

Similarly, the f_g is updated by f_l by the Global-to-Local Fusion. Specifically, a filter $F^{gl}(\cdot)$ is learned to update the global neighbor pixels (denoted as $G_{(i,j)}^{n^2}$) by the corresponding center pixel $L_{(i,j)}$ in f_l. The weight of $F^{gl}(\cdot)$ is calculated as:

$$\boldsymbol{\omega}_{(i,j)}^{1} = softmax\left(\sum\nolimits_{c}(L_{(i,j)} \otimes G_{(i,j)}^{n^2})\right) \tag{3}$$

Thus, the updated $G_{(i,j)}^{n^2}{}'$ can be obtained by:

$$G_{(i,j)}^{n^2}{}' = F^{gl}\left[G_{(i,j)}^{n^2}\right] = \sum\nolimits_{n^2}(G_{(i,j)}^{n^2} \otimes \boldsymbol{\omega}_{(i,j)}^{1}) \tag{4}$$

Therefore, we can get the fused information f_g' by updating all pixels in f_g. Finally, the fused features f_l' and f_g' are further added to generate the final fused features f_m:

$$f_m = f_l' \oplus f_g' \tag{5}$$

where $\oplus$ represents the concatenation along the channel dimension.

In simple terms, the MMF module utilizes the weighted operation in local areas and the softmax-based homologous attention mechanism to fuse features, which effectively captures the correlations between the local and global feature maps. Therefore, the information extracted from both the U-Net and Transformer branches can be fully exchanged and fused, thereby improving the feature representation.

B. TGCN Subnetwork

The topological interaction of different landmarks is important to provide reliable prior information for improving detection performance. However, existing landmark detection-based algorithms for DDH diagnosis ignore the topological information hidden in hip BUS images. Since each hip BUS image has been annotated with a class label (DDH patient or normal subject), a TGCN subnetwork is then proposed to effectively learn topology-aware graph representations by leveraging the image-level class labels as the weakly supervision information. This weakly supervised learning can further refine the generated heatmaps, because the label information can implicitly provide an additional constraint to correct the detected landmarks.

1) Landmark Topological Relations

As shown in Fig. 4(a), we model three groups of topological relationships from six hip landmarks inspired by the Graf's method [4]. That is, three critical lines of the related structures are formed by L_1 and L_2, L_3 and L_4, and L_5 and L_6, respectively, in Fig. 1(c). In order to make full use of this valuable topological information, a graph is then constructed for graph representation learning.

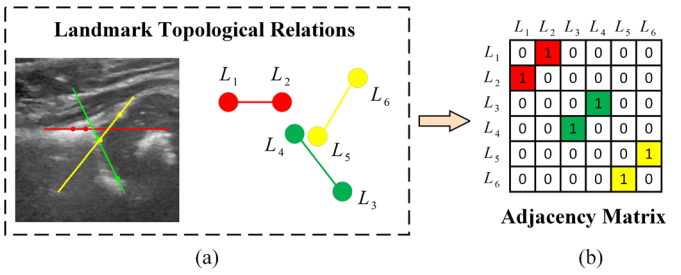

Fig. 4. The principle of the adjacency matrix construction. (a) Three groups of topological relations among hip landmarks; (b) Constructed adjacency matrix from the topological relations.

A graph is denoted as $G = (V, E)$, where V and E represent the nodes and a set of edges in the graph, respectively. Since each heatmap generated by the ICF subnetwork represents a corresponding landmark, we take each heatmap as a node. Thus, a graph can be denoted as a feature matrix $G_f \in \mathbb{R}^{k\times d}$, which has k nodes, and each node has a d-dimensional feature vector ($d = h \times w$ represents the size of each heatmap). In this work, we set $k = 6$, since our detection task will extract 6 landmarks.

2) Adjacency Matrix Construction

As shown in Fig. 4(b), we construct the adjacency matrix $A_{i,j} \in \mathbb{R}^{k\times k}$ based on the above mentioned three groups of topological relations. The adjacency matrix can be denoted as follows:

$$A_{i,j} = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & otherwise \end{cases} \tag{6}$$

where (v_i, v_j) is an edge between vertex v_i and v_j. For example, since L_1 and L_2 are collinear, we define $A_{1,2} = A_{2,1} = 1$. Thus, six edges of the graph can be constructed from the three groups of hip landmark topological relations.

In this way, the valuable topology information is then embedded into an adjacency matrix for further learning graph representations. Although the adjacency matrix is simple, it effectively represents the spatial relationships among the six landmarks. After modeling the additional spatial constraint relations, the subsequent GCN can well learn the graph representation with the help of image-level class labels in the weakly supervised learning paradigm, thus effectively enhancing the landmark detection performance.

After obtaining $G_f \in \mathbb{R}^{k\times d}$ and $A_{i,j} \in \mathbb{R}^{k\times k}$, they are fed into a multi-layer GCN [16]. The operation is given by

$$G_f{}' = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}G_f\boldsymbol{W}^{(l)}\right) \tag{7}$$

where $G_f{}' \in \mathbb{R}^{k\times d}$, $\sigma(\cdot)$ denotes an activation function, $\tilde{D}$ is the degree matrix of $\tilde{A}$, $\tilde{A} = A_{i,j} + I_N$, I_N is the identify matrix, and $\boldsymbol{W}^{(l)}$ is a trainable weight matrix. The graph representations are then fed into the liner projections to generate the final output:

$$y_{GCN} = (G_f{}'\boldsymbol{W}_0)\boldsymbol{W}_1 \tag{8}$$

where $\boldsymbol{W}_0 \in \mathbb{R}^{d_m\times d}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{d_c\times d_m}$ are the weight matrix of linear projection, and d_m and d_c represent the middle dimensionality and final class number, respectively.

Through above operations, the topological information in the hip ultrasound images can be effectively embedded into the GCN layers. That is, the graph representations contain valuable topological information, which are essential for improving the detection performance.

C. MSC Strategy

Existing landmark detection-based models mainly hypothesized that the landmarks are independent, and thus they

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3559383

only focus on regressing the coordinates or heatmaps of the landmarks, but ignore the spatial relationships among landmarks. In fact, each hip landmark is constrained by neighboring points in the spatial domain. To this end, inspired by [14], we propose a novel MSC strategy that provides additionally spatial constraint on both the horizontal and vertical axes for each landmark, thereby further enhancing the hip landmark detection accuracy.

In the MSC strategy, each hip landmark is constrained by the former one and the latter one. For example, the L_2 is associated with L_1 and L_3. Specifically, the supervision vector is defined to constrain each landmark on both horizontal and vertical axes by

$$\begin{cases} e_i = (x_{i+1} - x_i, y_{i+1} - y_i), & i = 1,2,\cdots,5 \\ e_6 = (x_1 - x_6, y_1 - y_6), & otherwise \end{cases} \quad (9)$$

where e_i represents the supervision vector of each hip landmark, and x_i and y_i are the coordinates of each landmark on x and y axes, respectively. Therefore, the ground truth mutual-supervision vector can be denoted as

$$v_m = [e_1, e_2, e_3, e_4, e_5, e_6] \quad (10)$$

In addition, the predicted mutual-supervision vector is then generated in ICF subnetwork. As shown in Fig. 2 (d), the fully fused and exchanged features f_m (according to Eq. (5)) are first fed into a global average pooling layer to generate the vectors. The subsequent reshape layer is used to reshape the vectors into the predicted mutual-supervision vector $v_m{}'$ as

$$v_m{}' = Reshape\big(GAP(f_m)\big) \quad (11)$$

Then, both the v_m and $v_m{}'$ are final fed into the landmark mutual supervision constraint loss L_{vector} that is evaluated by smooth L1 loss [32].

Therefore, each hip landmark is constrained by the nearest two landmarks on both the horizontal and vertical axes, respectively. That is, the MSC strategy can effectively model the spatial constraint relationships among neighbor landmarks, so as to improve the detection performance.

D. Loss Function

As shown in Fig. 2, the TGCN-ICF is trained by three loss functions, including the $L_{landmark}$, L_{vector}, and $L_{classify}$, since various recent works indicated that using a joint loss is superior to using a single loss [33], [34]. Specifically, the $L_{landmark}$ is calculated by the following Mean Square Error (MSE) loss [35] between the ground truth heatmaps and predicted heatmaps

$$L_{landmark} = \frac{1}{N}\sum_{i=1}^{N}(h^{pred} - h^{gt})^2 \quad (12)$$

where N is the number of total landmarks, h^{pred} denotes the heatmaps that generated by the ICF subnetwork, and h^{gt} represents the ground truth heatmaps.

The L_{vector} is evaluated by a smooth L1 loss [32] between the predicted and ground truth mutual-supervision vectors

$$L_{vector} = \begin{cases} 0.5(v_m{}' - v_m)^2, & if\ |v_m{}' - v_m| < 1 \\ |v_m{}' - v_m| - 0.5, & otherwise \end{cases} \quad (13)$$

where $v_m{}'$ and v_m represent the predicted mutual-supervision vector (as shown in Eq. (11)) and the ground truth mutual-supervision vector (as shown in Eq. (10)), respectively.

The $L_{classify}$ is evaluated by a Binary Cross Entropy (BCE) loss [36] between the ground truth labels and predicted classes, which is calculated by

$$L_{classify} = -\frac{1}{M}\sum_{i=1}^{M}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (14)$$

where M is the number of total hip ultrasound images, y_i is the class of $i\text{-}th$ hip image (0 represents the abnormal, and 1 represents the normal), $p_i \epsilon [0,1]$ denotes the predicted value of $i\text{-}th$ sample.

The final loss function L is calculated by

$$L = L_{landmark} + \lambda * L_{vector} + \mu * L_{classify} \quad (15)$$

where λ and μ are the hyperparameters to adjust the proportion of the three losses. By combining the above three loss functions, the proposed TGCN-ICF can be effectively trained for detecting the hip landmarks within BUS images.

Algorithm 1 provides the pseudocode of the training process of our model.

Algorithm 1 TGCN-ICF Training.

Input: Images $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$; Adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{k\times k}$.
Output: Predicted heatmaps $\boldsymbol{H}^i \in \mathbb{R}^{C\times H\times W}$; Predicted mutual-supervision vector $\boldsymbol{V}^i \in \mathbb{R}^{k\times 2}$; Predicted class $\boldsymbol{Y}^i \in \mathbb{R}^{l\times 1}$.

```
for each epoch do
    for each batch do
        X^I, X^P ← Data Processing (X)
        f_l ∈ R^{h×w×c} ← U-Net (X^I)
        f_g ∈ R^{h×w×c} ← Transformer (X^P)
        f_m ∈ R^{h×w×2c} ← MMF (f_l, f_g)
        H^i ← Detection Head (f_m)
        V^i ← MSC (f_m)
        Y^i ← TGCN Subnetwork (H^i, A)
    end for
end for
return H^i, V^i, Y^i
```

IV. Experiments

A. Datasets

Two real-world BUS DDH datasets were utilized to evaluate the effectiveness of the proposed TGCN-ICF, which were acquired from the Shanghai Children's Medical Center (SCMC DDH Dataset) and the Anhui Provincial Children's Hospital (APCH DDH Dataset), respectively.

The SCMC DDH Dataset consists of 700 hip ultrasound images from 413 infants, which was collected between June 2022 and October 2023. These images were scanned by two ultrasound imaging devices, namely LOGIO E9 (GE HealthCare, Milwaukee, WI) and SIEMENS OXANA 2 (SIEMENS, Chicago, IL, USA). Moreover, there were three sizes of image resolution, including 368×390, 440×480, and 480×480 pixels. All landmarks were marked by two experienced sonologists. This study was approved by the Research Ethics Board of Shanghai Children's Medical Center (No. SCMCIRB-K2023027-1), and informed consent was signed by all guardians of the infants.

The APCH DDH Dataset includes 1769 hip ultrasound

images, which were scanned by a Philips EPIQ 5 ultrasound system between December 2018 to November 2019 [7]. All images had the same resolution of 445×715 pixels. In addition, the landmarks were labeled and cross-validated by four professional sonologists, who have engaged in DDH diagnosis for more than five years.

B. Experimental Settings

1) Comparison Experiment

To evaluate the performance of the proposed TGCN-ICF, we compared it with the following classical and state-of-the-art (SOTA) landmark detection algorithms on both DDH datasets:

1) U-Net [37]: The classical U-Net model was applied for hip landmark detection.
2) DM-ResNet [7]: It was a specially proposed model for the hip landmark detection task in ultrasound images, which adopted a simple ResNet as the backbone with a novel dependency mining module to enhance feature representation for improving detection accuracy.
3) TransUNet [38]: It was a representative CNN-Transformer hybrid model, which adopted Transformer as strong encoders and employed U-Net to recover localized spatial information that enhanced details.
4) Conformer [13]: It was the original Conformer model but with the U-Net instead of CNN branch, which was compared as a baseline in this work.
5) FAT-Net [39]: It was a representative dual-branch network that utilized the CNN and Transformer as a dual encoder with three feature adaptation modules for fusing features.
6) FARNet [40]: This model was a novel encoder-decoder architecture for anatomic landmark detection that fused multi-scale features from the encoder to achieve high resolution heatmap regression.
7) DA-TransUNet [41]: It was a SOTA U-shape architecture, which utilized the Transformer and dual attention blocks to integrate both global and local features together with the image-specific positional and channel features.
8) SCUNet++ [42]: It was another SOTA network with multiple fused dense skip connections between the encoder and decoder, which aimed to fuse features of different scales to enhance feature representation.
9) AAU-Net [43]: It was an effective U-Net variant for ultrasound image segmentation, which designed a hybrid adaptive attention module to enhance feature representation in both the channel and space dimensions.
10) C-Net [44]: This model was a novel cascaded convolutional neural network that incorporated a bidirectional attention guidance network to capture the context between global and local features.
11) NU-Net [45]: It was an unpretentious nested U-Net for segmenting breast tumors in ultrasound images, which utilized U-Nets with different depths and shared weights to improve feature representation.
12) ESKNet [46]: This model employed an enhanced selective kernel convolution module to construct a novel deep supervised U-Net for adaptively capturing features from the channel and spatial dimensions.
13) IT-UNet [8]: It was a novel hip landmark detection algorithm, which integrated the involution operation into Transformer to capture both spatial and long-range information for hip landmark detection.

2) Ablation Study

We also conducted an ablation experiment on the SCMC DDH Dataset to compare the developed MMF module with the following fusion strategies:

1) Addition [47]: This variant used the simple addition strategy to fuse the features extracted from the CNN and Transformer branches of Conformer model.
2) Concatenation [48]: This variant employed the conventional concatenation strategy to fuse the features of U-Net and Transformer branches in Conformer.
3) BiFusion [28]: This variant utilized the BiFusion module from TransFuse model to fuse features, which incorporated both self-attention and multi-modal fusion mechanisms to efficiently fuse the multi-level features from both branches.
4) Attentional Feature Fusion (AFF) [29]: This variant employed the attentional feature fusion module to fuse the two branches' features.
5) Fusion [49]: This variant utilized the simple calculations (e.g., multiplication and addition) and convolutional operations to fuse the features from the two branches.
6) Collection Information Module (CIM) [50]: This variant employed a novel collection information module with excellent learning and generalization abilities to fuse the features.

In addition, we conducted another ablation experiment on the SCMC DDH Dataset to further verify the proposed TGCN subnetwork and the MSC strategy:

1) Conformer with MMF (Conformer-MMF): This variant only adopted the MMF module within Conformer model, without the proposed TGCN subnetwork and MSC strategy, which was also served as the baseline in this ablation experiment.
2) TGCN-ICF without TGCN (TGCN-ICF w/o TGCN): This variant removed the proposed TGCN subnetwork in the proposed TGCN-ICF, and then directly applied the improved Conformer subnetwork for detecting hip landmarks.
3) TGCN-ICF without MSC (TGCN-IC w/o MSC): This variant removed the proposed MSC strategy in the proposed TGCN-ICF.

C. Evaluation Metrics

We performed the five-fold cross-validation strategy, which was the same split in [8], to evaluate the effectiveness of all algorithms. All results were given in the format of mean ± SD (standard deviation). The mean radial error (MRE) and successful detection rate (SDR) were commonly adopted as the two evaluation indices in landmark detection [8]. Specifically, the MRE represents the mean radial error between the predicted and ground truth landmark, which is defined as:

$$MRE_n = \left| L_n^p - L_n^g \right|^2 \tag{16}$$

$$MRE = \frac{1}{N} \sum_{n=1}^{N} MRE_n \tag{17}$$

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3559383

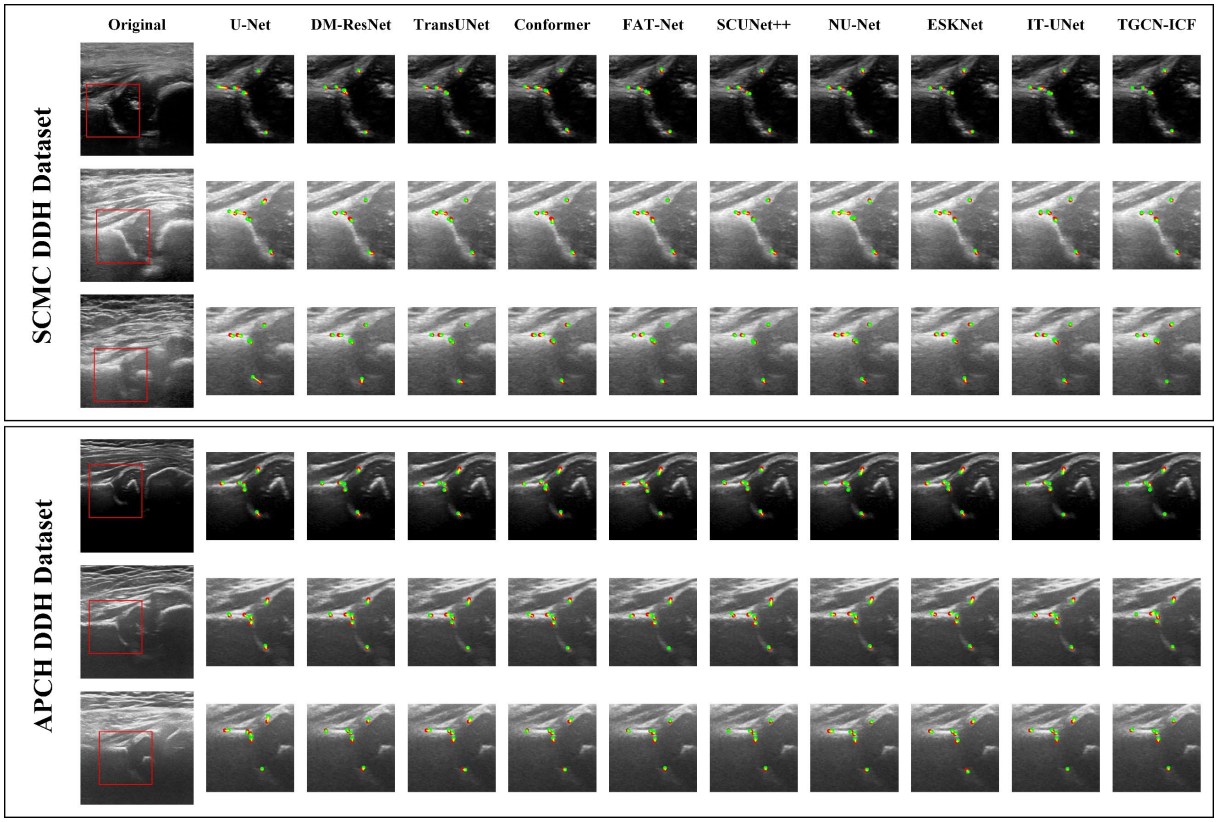

Fig. 5. Visualization examples by several representative landmark detection algorithms on two BUS DDH datasets.

where $L_n^p \in (x_n^p, y_n^p)$ and $L_n^g \in (x_n^g, y_n^g)$ represent the $n\text{-}th$ predicted and ground truth landmarks, respectively. The MRE_n represents the $n\text{-}th$ landmark's detection error, and the MRE denotes the average radial errors of all hip landmarks.

The SDR is the metrics to evaluate the distribution of MRE, which is calculated by

$$SDR_{dist} = \frac{\#\{m: MRE_n \le dist\}}{M} \times 100\% \qquad (18)$$

where # is the count symbol, m represents the number of landmarks that the mean radial error is less than $dist$, $dist$ denotes the scope of successful detection, and M is the total number of landmarks in the hip ultrasound image. In this work, we set the $dist$ into 0.5mm, 1.0mm, and 1.5mm, respectively.

TABLE I
QUANTITATIVE RESULTS OF DIFFERENT ALGORITHMS FOR HIP LANDMARK DETECTION ON THE SCMC DDH DATASET

Algorithms	MRE (mm) ↓	SDR (%) ↑		
		0.5mm	1.0mm	1.5mm
U-Net	0.4923±0.0255	67.12±1.62	92.48±0.83	97.14±0.71
DM-ResNet	0.4861±0.0262	68.64±2.13	91.14±1.41	96.29±0.59
TransUNet	0.5056±0.0258	67.57±1.25	91.79±0.98	96.45±0.62
Conformer	0.4828±0.0136	68.88±1.87	92.45±0.97	97.26±0.63
FAT-Net	0.4670±0.0219	69.41±0.60	93.17±0.76	97.45±0.76
FARNet	0.4706±0.0172	68.29±0.65	92.62±0.89	97.12±0.71
DA-TransUNet	0.4682±0.0213	69.60±0.92	92.17±0.61	96.71±0.84
SCUNet++	0.4742±0.0216	69.43±2.82	92.74±1.49	96.52±0.49
AAU-Net	0.4732±0.0133	68.86±1.34	92.10±0.69	96.75±0.74
C-Net	0.4663±0.0180	69.10±1.03	93.02±1.24	96.94±1.07
NU-Net	0.4604±0.0087	69.45±1.78	93.11±0.62	97.07±0.56
ESKNet	0.4549±0.0139	70.25±0.88	92.94±0.23	97.38±0.68
IT-UNet	0.4494±0.0155	71.19±1.76	93.45±1.07	97.31±0.56
TGCN-ICF	**0.4349±0.0148**	**72.62±1.37**	**94.86±0.44**	**98.45±0.78**

D. Implementation Details

In our implementations, the input ultrasound images were resized to 256×256. Meanwhile, the Adam optimizer was used for network optimization with an initial leaning rate of 1e-4, and the TGCN-ICF was trained for 300 epochs with a batch size of 2. Moreover, we set the hyperparameter σ to 10, which determined the Gaussian distribution when generating the ground truth heatmap of each hip landmark. The balance loss factors λ and μ were set to 1e-4 and 1e-5, respectively. All the landmark detection algorithms were implemented by PyTorch with two GTX 3090 GPUs.

V. EXPERIMENTAL DETAILS

A. Results of Comparison Experiment

Fig. 5 shows the visualization results of landmark detection from hip BUS images by different algorithms on both the SCMC and APCH DDH Dataset. The left red boxes are the areas of critical anatomical structures. Moreover, the red dots represent the ground truth landmarks, the green dots denote the predicted results, and the yellow lines between the red dots and green dots show the detected errors. It can be found that the predicted landmarks are closer to the ground truth landmarks by the proposed TGCN-ICF, which indicate that our detection algorithm achieves the best detection performance. It is worth noting that despite the variations across different imaging devices and medical institutions, the TGCN-ICF still achieves

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2025.3559383

the best detection accuracy. These superior visualization results demonstrate the effectiveness and robustness of the TGCN-ICF.

Table I shows the comparison results on the SCMC DDH Dataset. It is observed that the proposed TGCN-ICF outperforms all the compared algorithms on both metrics, with the best MRE of 0.4349±0.0148mm and SDRs of 72.62±1.37% (0.5mm), 94.88±0.44% (1.0mm), and 98.45±0.78% (1.5mm), respectively. Compared to the specially designed landmark detection algorithms (DM-ResNet, FARNet, IT-UNet), the TGCN-ICF declines at least 0.0145mm (approximately 3.23%) on MRE, and also improves at least 1.43%, 1.41%, and 1.00% on SDR of 0.5mm, 1.0mm, and 1.5mm, respectively. Moreover, the TGCN-ICF also surpasses all the representative CNN-Transformer hybrid models, including TransUNet, Conformer, FAT-Net, and DA-TransUNet, on both MRE and SDRs. It also outperforms the specially designed ultrasound image analysis algorithms, including AAU-Net, C-Net, NU-Net, and ESKNet. In addition, Fig. 6 shows the comparison of detection performance for each hip landmark with several representative algorithms. The proposed TGCN-ICF achieves the best detection accuracy for almost all landmarks, with the exception of the landmark 6. All these superior results demonstrate the effectiveness of TGCN-ICF in detecting hip landmarks from BUS images.

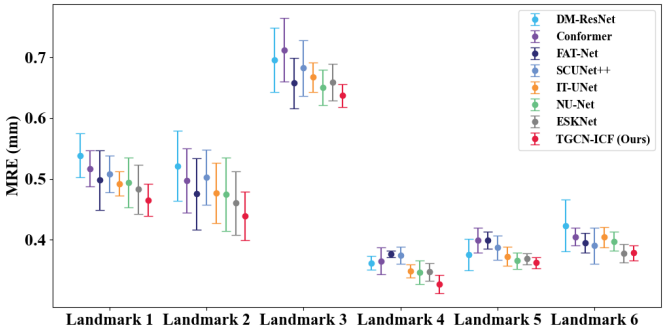

Fig. 6. Comparison of MRE results for each hip landmark with several representative algorithms on SCMC DDH Dataset.

Table II further shows the quantitative results of different algorithms on the APCH DDH Dataset. It can be fund that the quantitative results exhibit a similar trend to those in Table I. Our TGCN-ICF again achieves the best detection performance on the MRE and three SDR metrics. Specifically, it obtains the best MRE of 0.4133±0.0236mm, which decreases at least 0.0149mm (about 3.48%) in comparison to other algorithms. Additionally, the TGCN-ICF also gets the highest scores on the SDR at 0.5mm, 1.0mm and 1.5mm, with values of 73.09±1.03%, 94.80±0.10%, and 98.49±0.20%, respectively. Moreover, Fig. 7 illustrates the MRE results of each hip landmark with some representative algorithms. It can be found that the TGCN-ICF still outperforms all the compared algorithms for each point in BUS images.

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT ALGORITHMS FOR HIP LANDMARK DETECTION ON THE APCH DDH DATASET

Algorithms	MRE (mm) ↓	SDR (%) ↑		
		0.5mm	1.0mm	1.5mm
U-Net	0.4661±0.0266	67.72±0.92	93.56±0.35	98.10±0.21
DM-ResNet	0.4567±0.0253	70.24±0.62	93.16±0.39	97.69±0.41
TransUNet	0.4685±0.0261	69.25±1.32	92.42±0.31	97.44±0.19
Conformer	0.4442±0.0240	70.86±0.45	94.01±0.21	98.20±0.11
FAT-Net	0.4388±0.0215	71.07±1.45	94.23±0.17	98.19±0.26
FARNet	0.4470±0.0193	70.22±1.18	93.85±0.33	98.29±0.18
DA-TransUNet	0.4426±0.0270	71.13±1.37	93.77±0.46	98.12±0.18
SCUNet++	0.4406±0.0243	70.08±0.66	93.80±0.33	98.09±0.25
AAU-Net	0.4459±0.0188	69.39±1.73	93.37±0.80	98.05±0.24
C-Net	0.4382±0.0132	70.93±1.98	93.59±0.14	97.86±0.19
NU-Net	0.4357±0.0155	70.84±1.97	93.50±0.50	97.94±0.22
ESKNet	0.4307±0.0164	71.49±2.24	94.03±0.18	98.16±0.17
IT-UNet	0.4282±0.0206	72.19±1.60	94.25±0.43	98.14±0.24
TGCN-ICF	**0.4133±0.0236**	**73.09±1.03**	**94.80±0.10**	**98.49±0.20**

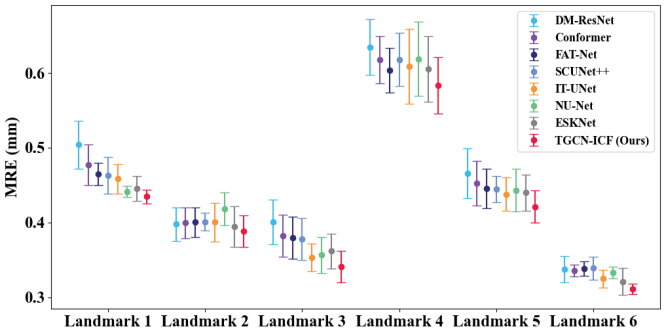

Fig. 7. Comparison of MRE results for each hip landmark with several representative algorithms on APCH DDH Dataset.

B. Results of Ablation Study

Fig. 8 illustrates the visual comparison of two ablation studies on the SCMC DDH Dataset. Specifically, compared with other feature fusion variants, the developed MMF variant obtains the most superior visual detection performance. This observation suggests the effectiveness of the MMF. Moreover, the variants TGCN-ICF w/o TGCN and TGCN-ICF w/o MSC show visualized decline compared to the TGCN-ICF, indicating the importance of the proposed TGCN and MSC.

Table III gives the quantitative ablation results of different fusion methods on the SCMC DDH Dataset. The developed MMF variant achieves the best detection accuracy, with the MRE value of 0.4682±0.0254mm and three SDR values of 69.60±0.90% (0.5mm), 93.31±1.19% (1.0mm), and 97.52±0.73% (1.5mm), respectively. In comparison to other conventional and effective fusion strategies, it reduces the MRE by at least 0.0061mm (approximately 1.29%). Additionally, it also demonstrates improvements at least 0.91% on SDR at thresholds of 0.5mm. Compared to the more recent fusion

TABLE III
ABLATION STUDY OF DIFFERENT FEATURE FUSION METHODS ON THE SCMC DDH DATASET

Methods	MRE (mm) ↓	SDR (%) ↑		
		0.5mm	1.0mm	1.5mm
Addition	0.4884±0.0288	67.93±2.42	92.54±0.65	97.14±0.61
Concatenation	0.4828±0.0136	68.58±1.87	92.45±0.97	97.26±0.63
BiFusion	0.4743±0.0278	68.52±1.63	92.48±1.05	97.24±0.71
AFF	0.4854±0.0204	67.98±1.76	92.31±0.75	97.25±0.59
Fusion	0.4806±0.0250	67.97±2.79	92.47±1.39	97.20±1.20
CIM	0.4783±0.0202	68.69±1.49	92.66±0.45	97.31±0.56
MMF (Ours)	**0.4682±0.0254**	**69.60±0.90**	**93.31±1.19**	**97.52±0.73**

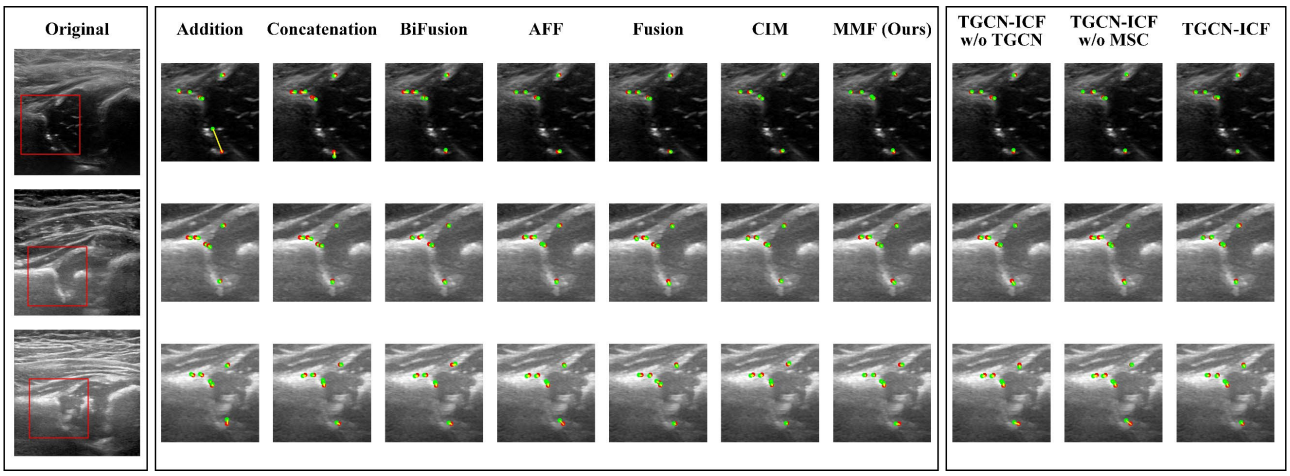

Fig. 8. Visualization results of two ablation experiments on SCMC DDH Dataset.

module CIM, the developed MMF variant also achieves better detection performance on both MRE and SDRs. These quantitative results suggest the effectiveness of the MMF in deeply fusing and exchanging information extracted from U-Net and Transformer branches.

Table IV further shows the quantitative results of ablation study to evaluate the proposed TGCN and MSC. It is notable that after removing the TGCN from the TGCN-ICF, the variant TGCN-ICF w/o TGCN exhibits an increase of 0.0232mm (about 5.06%) on MRE, and a reduction of 1.79% on SDR at 0.5mm. Additionally, the variant TGCN-ICF w/o MSC that still has TGCN declines 0.0118mm (approximately 2.52%) on MRE, and improves 1.90% on SDR (0.5mm) compared to Conformer-MMF (Baseline in this ablation experiment). It demonstrates the effectiveness of TGCN subnetwork to learn topological graph representations with the guidance of class labels. Moreover, the variant TGCN-ICF w/o MSC increases 0.0215mm on MRE, approximately 4.71% compared to the TGCN-ICF. In comparison to the Conformer-MMF, the variant TGCN-ICF w/o TGCN that still has MSC shows a reduction of 0.0101mm (about 2.16%) on MRE. These results prove the importance of MSC strategy to exploit the constraint relationships among different hip landmarks.

TABLE IV
ABLATION STUDY OF TGCN AND MSC ON THE SCMC DDH DATASET

Methods	MRE (mm) ↓	SDR (%) ↑		
		0.5mm	1.0mm	1.5mm
Conformer-MMF	0.4682±0.0254	69.60±0.90	93.31±1.19	97.52±0.73
TGCN-ICF w/o TGCN	0.4581±0.0219	70.83±1.05	94.06±1.00	97.80±0.87
TGCN-ICF w/o MSC	0.4564±0.0221	71.50±0.80	94.23±0.87	97.87±0.83
TGCN-ICF	**0.4349±0.0148**	**72.62±1.37**	**94.86±0.44**	**98.45±0.78**

C. Computational Complexity

As shown in Fig. 9, we further present the comparison of the computational complexity for different models. The horizontal axis in the figure represents the model parameters, and the vertical axis denotes the floating point operations (FLOPs). The proposed TGCN-ICF has a parameter count of 177.074M and a computational cost of 65.127G FLOPs. Compared to other algorithms, the TGCN-ICF increases the network parameters but without significantly increasing the computational cost. Thus, the TGCN-ICF achieves superior detection performance, indicating a better trade-off between the landmark detection accuracy and the computational complexity. On the other hand, the proposed TGCN-ICF does not significantly increase either the parameters or FLOPs compared with the backbone Conformer, but achieves much higher landmark detection accuracy than Conformer, as shown in Table I and Table II. It demonstrates the effectiveness and efficiency of the proposed TGCN subnetwork, MMF module, and MSC strategy in TGCN-ICF.

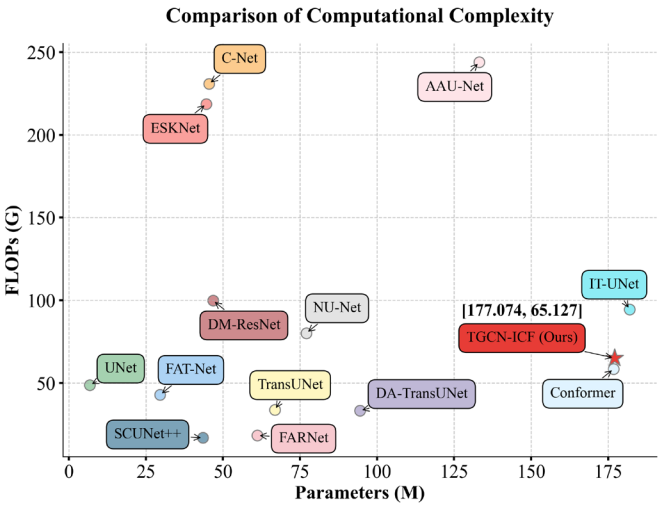

Fig. 9. Comparison of the computational complexity for different models.

VI. DISCUSSION

In this work, we propose a novel TGCN-ICF to detect six hip landmarks within BUS images, which then can be further used to calculate the α and β angles for diagnosis of DDH based on the Graf's method. The experimental results on two real-world DDH datasets demonstrate the effectiveness of the proposed TGCN-ICF, indicating its potential applications in the CAD of DDH for clinical practice.

It is worth noting that the hip ultrasound images are affected by speckle noise, making it a challenging task to accurately detect the critical landmarks. According to the Graf's method

in clinical practice for diagnosis of DDH [4], there exist special topological relations among the six hip landmarks. Besides, the class labels of each ultrasound image can also provide additional supplementary information to refine the detection. However, the previous landmark detection based works for DDH diagnosis do not pay attention to these important prior information. To this end, we model three groups of topological relations among the six hip landmarks by constructing a special adjacency matrix, and then innovatively design a TGCN subnetwork with the guidance of class labels. The experimental results indicate that the additional TGCN subnetwork can effectively learn graph representations to refine the generated heatmaps from the ICF subnetwork.

On the other hand, most of the landmark detection approaches generally hypothesize that each point is independent. In fact, besides the inherent topological relations among different hip landmarks, these points also exhibit positional relations in the spatial domain of hip ultrasound images. Although previous works have attempted to construct spatial relations by defining edges among landmarks [14], it is still limited in supervising every point. To this end, we propose a new MSC strategy to provide spatial constraint for each hip landmark. In MSC, all six hip landmarks can form supervisions and constraints in positional relationships. Both quantitative and visualized results of the ablation experiment demonstrate the effectiveness of the proposed MSC.

Existing feature fusion methods can be approximately divided into three categories: simple fusion (e.g., Addition [47], Concatenation [48]), convolution-based fusion (e.g., AFF [29]), and attention-based fusion (e.g., BiFusion [28]). Instead of these feature fusion approaches, we design a novel modulation-based fusion module, namely MMF, inspired by [31]. The MMF initially modulates and optimizes each branch's information by another one, and subsequently fuses them together. In this way, the features that extracted by CNN and Transformer branches can be fully exchanged and fused, so as to enhance the feature representation performance in the proposed TGCN-ICF.

According to the experimental results presented in Table I and II, it is evident that the proposed TGCN-ICF outperforms all the compared algorithms on both the MRE and SDRs metrics in the hip landmark detection task. Considering the complexity of clinical scenarios, we evaluate the TGCN-ICF on two real-world DDH datasets acquired from two hospitals with three different ultrasound devices. The visualization results shown in Fig. 5 further illustrate that the TGCN-ICF achieves the superior detection performance. Moreover, as shown in Fig. 9, the proposed TGCN-ICF exhibits the medium FLOPs compared to other comparison algorithms, which indicates its strong performance in model inference. All these experimental results demonstrate the effectiveness and efficiency of the proposed TGCN-ICF. After accurately detecting the anatomical hip points in ultrasound images, the detected landmarks serve as the key points to form three critical lines, which are then used to calculate the α and β angles for DDH diagnosis. Thus, this CAD can assist sonologists to improve diagnosis accuracy, reduce workload and promote efficiency. That is, the proposed TGCN-ICF demonstrates its effectiveness in the hip landmark detection task, indicating its significant clinical potential for DDH diagnosis.

Although the proposed TGCN-ICF achieves superior performance to the compared algorithms, it still has room for improvement. For example, the TGCN-ICF is developed for detecting critical landmarks from static ultrasound images, and it cannot be directly applied to detect points from ultrasound videos. In fact, it is subjective and time-costing for sonologists to select an image as the standard plane of the infantile hip for angle measurement during the scanning process. Therefore, we will focus on developing the fast landmark detection models for real-time calculation of α and β angles from ultrasound videos in future, so as to make it really work for DDH diagnosis in clinical practice.

Moreover, the effectiveness of the proposed TGCN-ICF has not been investigated for detecting structures from other medical images, such as breast cancer or skin lesion detection, which are very challenging due to different imaging techniques and noise [51], [52]. In future work, we will further develop our TGCN-ICF model, adapting and applying it to medical image detection tasks, thereby enhancing its potential for broader application in clinical settings.

VII. Conclusion

In summary, we propose a novel weakly supervised TGCN-ICF algorithm for hip landmark detection from B-mode ultrasound images. Different from conventional heatmap regression-based approaches for landmark detection, we develop an additional TGCN subnetwork to explore the topological relations among different hip points for the refinement of the generated heatmaps. Moreover, a new MMF feature fusion module is designed in the ICF subnetwork, which aims to fully fusing and exchanging the information that extracted by the U-Net and Transformer branches. Meanwhile, we also propose a novel MSC strategy for providing spatial constraints of each detected landmark, so as to further enhance the detection accuracy. The experimental results on two real-world DDH datasets demonstrate the effectiveness and robustness of the TGCN-ICF, indicating its potentially clinical application.

References

[1] M. D. Sewell et al., "Developmental dysplasia of the hip," *Bmj.*, vol. 339, 2009.

[2] S. Sioutis et al., "Developmental dysplasia of the hip: a review," *J. Long-Term Eff. Med. Implants*, vol. 32, no. 3, pp. 39-56, 2022.

[3] D. Zhang et al., "Multi-frequency therapeutic ultrasound: A review," *Ultrason. Sonochem.*, 106608, 2023.

[4] R. Graf, "Fundamentals of sonographic diagnosis of infant hip dysplasia," *J. Pediatr Orthop.*, vol. 4, no. 6, pp. 735-740, 1984.

[5] R. Liu et al., "NHBS-Net: A feature fusion attention network for ultrasound neonatal hip bone segmentation," *IEEE Trans. Med. Imaging*, vol. 40, no. 12, pp. 3446-3458, 2021.

[6] A. Stamper et al., "Infant hip screening using multi-class ultrasound scan segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging*, pp. 1-4. 2023.

[7] J. Xu et al., "Hip landmark detection with dependency mining in ultrasound image," *IEEE Trans. Med. Imaging*, vol. 40, no. 12, pp. 3762-3774, 2021.

[8] T. Huang et al., "Involution Transformer based U-Net for Landmark Detection in Ultrasound Images for Diagnosis of Infantile DDH," *IEEE J. Biomed. Health Inform.,* 2024.

[9] J. Liu et al., "Speckle noise reduction for medical ultrasound images based on cycle-consistent generative adversarial network," *Biomed. Signal Process.*, vol. 86, 105150, 2023.

[10] S. S. Kshatri et al., "Convolutional neural network in medical image analysis: A review," *Arch. Comput. Methods Eng.*, vol. 30, no. 4, pp. 2793-2810, 2023.

[11] F. Shamshad et al., "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, 102802, 2023.

[12] T. Viriyasaranon et al., "Anatomical Landmark Detection Using a Multiresolution Learning Approach with a Hybrid Transformer-CNN Model," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2023, pp. 433-443.

[13] Z. Peng et al., "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 367-376.

[14] W. Liu et al., "Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2020, pp. 670-679.

[15] L. Zhang et al., "Graph neural networks for image-guided disease diagnosis: A review," *iRADIOLOGY*, vol. 1, no. 2, pp. 151-166, 2023.

[16] T. N. Kipf et al., "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[17] W. Ju et al., "A comprehensive survey on deep graph representation learning," *Neural Netw.*, 106207, 2024.

[18] T. Huang et al., "Topological GCN for Improving Detection of Hip Landmarks from B-Mode Ultrasound Images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2024, pp. 692-701.

[19] X. Hu et al., "Joint landmark and structure learning for automatic evaluation of developmental dysplasia of the hip," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 345-358, 2021.

[20] W. Chen et al., "TW-GAN: Topology and width aware GAN for retinal artery/vein classification," *Med. Image Anal.*, vol. 77, 102340, 2022.

[21] S. Gupta et al., "Learning topological interactions for multi-class medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 701-718.

[22] H. He et al., "Toposeg: Topology-aware nuclear instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21307-21316.

[23] S. Ding et al., "Multi-scale efficient graph-transformer for whole slide image classification," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 12, pp. 5926-5936. 2023.

[24] T. Zhang et al., "Topological structure and global features enhanced graph reasoning model for non-small cell lung cancer segmentation from CT," *Phys. Med. Biol.*, vol. 68, no. 2, 025007, 2023.

[25] Y. Wang et al., "HandGCNFormer: A Novel Topology-Aware Transformer Network for 3D Hand Pose Estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5675-5684.

[26] P. Shi et al., "Nextou: Efficient topology-aware u-net for medical image segmentation," 2023, arXiv preprint arXiv:2305.15911.

[27] R. Azad et al., "Advances in medical image analysis with vision transformers: a comprehensive review," *Med. Image Anal.*, vol. 91, 103000, 2023.

[28] Y. Zhang et al., "Transfuse: Fusing transformers and cnns for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 14-24.

[29] Y. Dai et al., "Attentional feature fusion," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3560-3569.

[30] F. Liu et al., "Dbmf: Dual branch multiscale feature fusion network for polyp segmentation," *Comput. Biol. Med.*, vol. 151, 106304, 2022.

[31] X. Dong et al., "Learning mutual modulation for self-supervised cross-modal super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1-18.

[32] R. Girshick, "Fast r-cnn," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1440-1448.

[33] E. Goceri, "Polyp segmentation using a hybrid vision transformer and a hybrid loss function," *J. Imaging Informat. Med.*, vol. 37, no. 2, pp. 851-863, 2024.

[34] E. Goceri, "GAN based augmentation using a hybrid loss function for dermoscopy images," *Artif. Intell. Rev.*, vol. 57, no. 9, 234, 2024.

[35] J. Ren et al, "Balanced mse for imbalanced visual regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7926-7935.

[36] T. Wu et al., "Adaptive spatial-bce loss for weakly supervised semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 199-216.

[37] O. Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234-241.

[38] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," in *Proc. Int. Conf. Mach. Learn. Workshop Interpretable Mach. Learn. Healthcare*, 2021, pp. 1-13.

[39] H. Wu et al., "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, 102327, 2022.

[40] Y. Ao et al., "Feature aggregation and refinement network for 2D anatomical landmark detection," *J. Digit. Imaging*, vol. 36, no. 2, pp. 547-561, 2023.

[41] G. Sun et al., "DA-TransUNet: Integrating Spatial and Channel Dual Attention with Transformer U-Net for Medical Image Segmentation," *Front. Bioeng. Biotechnol.*, vol. 12, 2024.

[42] Y. Chen et al., "Scunet++: Swin-unet and cnn bottleneck hybrid architecture with multi-fusion dense skip connection for pulmonary embolism ct image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7759-7767.

[43] G. Chen et al., "AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images," *IEEE Trans. Med. Imaging*, vol. 42, no. 5, pp. 1289-1300, 2022.

[44] G. Chen et al., "C-Net: Cascaded convolutional neural network with global guidance and refinement residuals for breast ultrasound images segmentation," *Comput. Methods Programs Biomed.*, vol. 225, 107086, 2022.

[45] G. Chen et al., "Rethinking the unpretentious U-net for medical ultrasound image segmentation," *Pattern Recognit.*, vol. 142, 109728, 2023.

[46] G. Chen et al., "ESKNet: An enhanced adaptive selection kernel convolution for ultrasound breast tumors segmentation," *Expert Syst. Appl.*, vol. 246, 123265, 2024.

[47] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.

[48] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818-2826.

[49] H. Zhu et al., "I can find you! boundary-guided separated attention network for camouflaged object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3608-3616.

[50] B. Xiao et al., "CTNet: Contrastive Transformer Network for Polyp Segmentation," *IEEE Trans. Cybernet.*, 2024.

[51] F. Z. Nakach et al., "A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification," *Artif. Intell. Rev.*, vol. 57, no. 12, 327, 2024.

[52] E. Goceri, "Automated skin cancer detection: where we are and the way to the future," in *Proc. IEEE Int. Conf. Telecommun. Signal Process.*, 2021, pp. 48-51.