Sensors and Materials, Vol. 38, No. 6 (2026) 3209–3224
MYU Tokyo

S & M 4498

Dual-stage Multi-attention Deep Learning Framework for X-ray-based Classification of Rotator Cuff Tears

Chen-Chou Hsieh,[1†] Shang-Lin Hsieh,[2†] Yen-Yu Chen,[3]
Hsien-Chu Wu,[3] Chwei-Shyong Tsai,[4*] Wei-Cheng Chang,[4]
Kuei-Wen Chen,[5] Yi-Cheng Yang,[6] and Chin-Jung Hsu[1]

[1]Department of Orthopedic Surgery, China Medical University Hospital,
No. 2, Yude Road, North District, Taichung 40447, Taiwan
[2]Department of Orthopedic Surgery, Taichung Municipal Geriatric Rehabilitation General Hospital,
No. 1141, Sec. 3, Taiyuan Rd., Beitun Dist., Taichung 406004, Taiwan
[3]Department of Artificial Intelligence and Computer Engineering, National Chin-Yi University of Technology,
No. 57, Sec. 2, Zhongshan Rd., Taiping Dist., Taichung 411030, Taiwan
[4]Department of Information Management, National Chung Hsing University,
145 Xingda Rd., South Dist., Taichung 40227, Taiwan
[5]Faculty of Engineering and Information Technology, The University of Melbourne,
Grattan Street, Parkville, Victoria 3010, Australia
[6]Faculty of Science, The University of Melbourne, Grattan Street, Parkville, Victoria 3010, Australia

(Received January 1, 2026; accepted May 28, 2026)

Keywords: rotator cuff tear, magnetic resonance imaging, X-ray, greater tuberosity sclerosis, deep learning

Rotator cuff tear (RCT) is a common cause of shoulder pain resulting from tendon or muscle injury. Although magnetic resonance imaging (MRI) is the clinical gold standard for RCT diagnosis, its high cost and limited accessibility often delay treatment. While conventional X-ray imaging sensors are widely available and cost-effective, they lack direct visualization of soft tissues. To maximize the diagnostic utility of radiographic sensor data, we propose a two-stage deep learning framework that leverages indirect indicators, particularly greater tuberosity sclerosis (GTS), to assess RCT severity. In the segmentation stage, a multi-attention–gated U-Net with full-scale skip connections accurately delineates GTS regions from the X-ray sensor images. In the classification stage, a dual-branch convolutional network integrates the acquired sensor data and GTS masks using spatial and channel attention mechanisms to classify RCTs into partial- or full-thickness tears. The segmentation model achieved a Dice coefficient of 0.835 and an accuracy of 0.998, outperforming several state-of-the-art methods. The proposed classification network reached an overall accuracy of 0.941, which surpasses those of previously reported MRI-based and X-ray-based approaches. This framework demonstrates how advanced computational technologies can significantly augment the diagnostic capabilities of standard X-ray sensing systems, enabling accurate and efficient RCT assessment, reducing reliance on MRI, and supporting timely clinical decision-making.

*Corresponding author: e-mail: ihunlee5@gmail.com
†These authors contributed equally to this work.
https://doi.org/10.18494/SAM6155

ISSN 0914-4935
https://myukk.org/

1. Introduction

Shoulder pain is the third most common musculoskeletal complaint encountered in primary care, with approximately half of the patients continuing to experience symptoms after six months of treatment.[1,2] Among its causes, rotator cuff tear (RCT) is particularly prevalent and represents a major source of disability in middle-aged and elderly populations. The rotator cuff, comprising four muscles and their tendons, stabilizes the humeral head within the glenoid cavity. The severity of RCTs ranges from mild partial-thickness tears causing minimal discomfort to large full-thickness tears that significantly impair shoulder function and are challenging to repair. As the tear progresses, muscle weakness and tendon retraction increase, reducing the success of conservative management.[3–5] Standard treatment strategies include rest, analgesics, and physiotherapy, while surgical repair is often required for severe or persistent cases.

Diagnosing RCT remains challenging because its clinical presentation overlaps with other shoulder pathologies such as adhesive capsulitis, tendinopathy, and labral lesions. Radiography is typically the first-line imaging modality owing to the low cost and widespread accessibility of X-ray sensors. Although conventional X-ray sensing systems cannot directly visualize soft tissue, the captured radiographic sensor data can reveal indirect indicators of RCT—most notably, greater tuberosity sclerosis (GTS)—which reflects subchondral bone reaction to chronic tendon traction.[6–9] Several studies have demonstrated significant associations between GTS and the presence or severity of RCT, with reported sensitivities exceeding 80% when analyzing these sensor-derived features in selected cohorts.[10–12] Additional radiographic findings, such as acromiohumeral interval narrowing and subchondral cysts, further support diagnosis based on X-ray sensor imaging.[13,14]

Conversely, magnetic resonance imaging (MRI) remains the gold standard for RCT evaluation, providing superior visualization of tear morphology, muscle atrophy, and fatty infiltration.[15,16] However, MRI is costly, time-consuming, and not universally available. Recent advances in AI and deep learning have shown promising potential for automating RCT detection and grading from both MRI and radiographic data.[17–21] Therefore, developing an X-ray-based AI system for RCT classification can facilitate earlier diagnosis, reduce unnecessary MRI examinations, and improve clinical efficiency in routine practice.

With the advancement of computer-aided diagnosis and AI, the accuracy and efficiency of medical image analysis have considerably improved. Deep neural networks are increasingly applied in computer-aided medical diagnosis, and many studies have utilized MRI for RCT classification and segmentation.[22] Although AI-assisted shoulder MRI aids clinical diagnosis, it does not shorten treatment time or reduce costs. Using AI to diagnose RCT from plain radiographs can substantially reduce MRI waiting times and unnecessary expenses.

In the field of medical image segmentation, the U-Net architecture[23] has demonstrated outstanding performance. Its symmetric encoder–decoder structure enables hierarchical feature extraction and precise spatial reconstruction. The skip connections between corresponding layers facilitate the transfer of contextual information, preserving fine anatomical details that are often lost in traditional convolutional pipelines. Building on this foundation, U-Net3+[24] introduced full-scale skip connections that connect multiple encoder and decoder layers

simultaneously. This enhancement allows the fusion of both high- and low-level semantic features, improving segmentation accuracy and edge delineation for complex anatomical structures.

The application of convolutional neural networks (CNNs) in medical image classification has considerably enhanced diagnostic accuracy and reliability, showing strong performance in disease detection and anomaly identification. Cho *et al.*[25] applied CNNs to diagnose RCT from shoulder X-ray images, demonstrating that convolution, subsampling, and dense layers are effective in extracting critical image features. Kim *et al.*[26] incorporated both image and unstructured data for binary classification, while the convolutional block attention module (CBAM)[27] further improved feature extraction by integrating channel and spatial attention mechanisms. Shim *et al.*[18] employed 3D CNNs for MRI-based RCT analysis, enabling richer spatial feature learning. Hashimoto *et al.*[28] used EfficientNet[29] with SENet to maintain accuracy while reducing computational cost, achieving refined RCT severity classification.

The primary contributions of this work are threefold as follows.

We propose the full-scale and multi-attention segmentation network (FSMA-SN), a novel segmentation network that leverages multi-attention gates and full-scale skip connections to precisely delineate GTS, providing a robust automated tool for identifying indirect radiographic signs of RCT.

We develop a dual GTS-Attention classification framework (Dual GTSA-CN) that integrates raw X-ray features with segmented pathological priors. This cross-modal guidance mechanism improves the differentiation between partial- and full-thickness tears, achieving performance superior to several state-of-the-art methods.

We demonstrate the clinical feasibility of an X-ray-only diagnostic pipeline, which offers a cost-effective and accessible alternative to MRI, potentially accelerating clinical decision-making in primary care settings.

The remainder of this paper is organized as follows. In Sect. 2, the details of the proposed rotator cuff tear segmentation and classification (RCTSC) framework, including the architectural design of the segmentation and classification modules, are given. The experimental results and comparative performance analysis are presented in Sect. 3. In Sect. 4, we discuss the clinical implications, limitations, and future directions, and Sect. 5 concludes the paper.

2. Data, Materials, and Methods

The proposed RCTSC framework is designed as a sequential sensing and diagnostic pipeline, as illustrated in Fig. 1. The architecture comprises two integrated stages: (1) FSMA-SN for GTS segmentation and (2) Dual GTSA-CN for tear classification.

The core interaction between these modules is defined by a dynamic spatial-weighting mechanism. Unlike conventional single-stage models, the RCTSC framework utilizes the binary mask generated by the X-ray sensor-based segmentation stage as a spatial prior. This mask provides topographic guidance to the classification stage, explicitly directing the attention mechanisms to focus on clinically relevant radiographic indicators (i.e., GTS) while suppressing noise from irrelevant anatomical structures. This coupling ensures that the classification

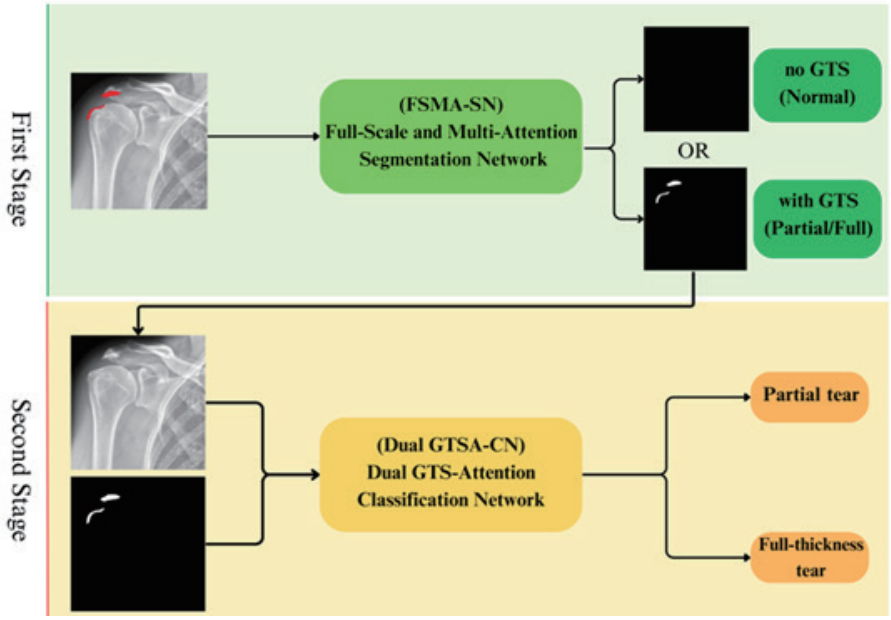

Fig. 1. (Color online) Framework of RCTSC.

decision is physically grounded in the pathological features extracted from the radiographic sensing data.

2.1 FSMA-SN

In the first stage, the FSMA-SN is introduced to improve lesion localization from the X-ray sensor images. Based on the U-Net architecture, this network incorporates multi-attention gates and full-scale skip connections. These enhancements allow the model to capture multiscale features and subtle bone density variations, ensuring high-fidelity delineation of the GTS region, which serves as the foundational input for the subsequent classification stage.

The indirect radiographic sign of GTS on shoulder X-ray images serves as an important indicator of RCT. However, manual interpretation remains time-consuming and labor-intensive in clinical practice. In the segmentation phase of RCTSC, we propose the FSMA-SN (Fig. 2), which enhances the full-scale skip connections of U-Net3+ to improve multiscale feature fusion. The network integrates multilevel features through multiple attention gates and consists of two main modules: the multi-attention gate and the full-scale skip connection.

2.2 Multi-attention gate

The design of FSMA-SN is specifically motivated by the challenging nature of detecting GTS in conventional radiographs. Unlike distinct anatomical structures, GTS often presents as subtle, heterogeneous increases in bone density, which lack sharp boundaries and can be easily confounded by overlapping shadows of the acromion or surrounding trabecular patterns. While the U-Net3+ architecture offers the advantage of full-scale skip connections to capture

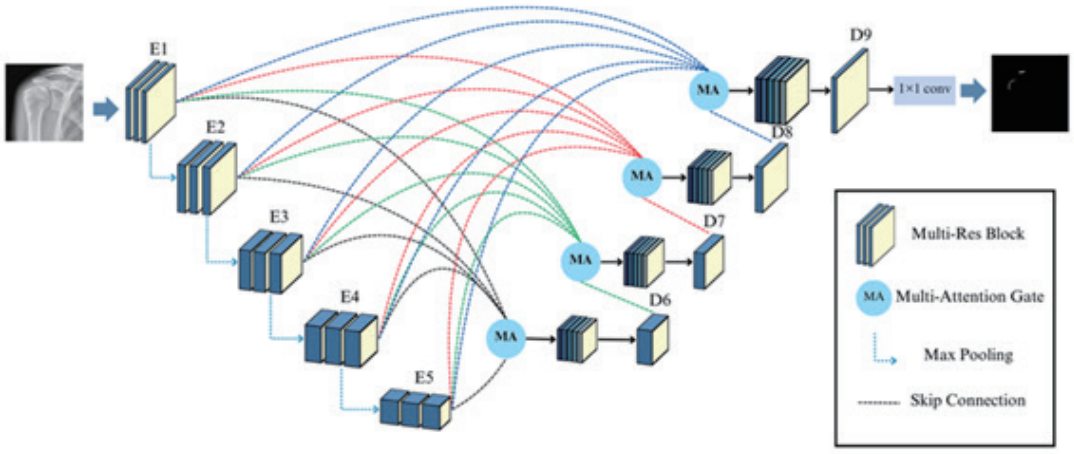

Fig. 2. (Color online) FSMA-SN architecture.

multiscale semantic information, it remains prone to incorporating irrelevant background noise in complex orthopedic images. To address this, we integrated a multi-attention gate mechanism. This theoretical enhancement allows the model to adaptively assign higher weights to salient osseous features while suppressing noninformative regions, thereby ensuring that the segmentation process remains robust against the high intraclass variability typical of sclerotic lesions.

To improve segmentation accuracy and quality, a multi-attention gate was applied at multiple levels to integrate features across layers (Fig. 3), enhancing the model's ability to capture both local details and global context. In addition, dilated convolutions were incorporated to enlarge the receptive field, enabling broader contextual information capture without increasing computational cost or the number of parameters.

(1) Input features X and Y from different levels are processed through dilated convolutions with different dilation rates, which expand the receptive field without additional parameters.
(2) Afterward, the outputs are passed through a 1 × 1 convolution for channel matching, resulting in Dx and Dy, which are then combined and activated by a ReLU function to suppress negative values while retaining positive features.
(3) A subsequent 1 × 1 convolution reduces the channel dimension to 1, and a sigmoid function normalizes the feature weights into attention coefficients.
(4) Finally, the input feature X is multiplied by the attention coefficient map to generate the output feature map, allowing the model to emphasize important features and suppress less relevant ones.

2.3 Full-scale skip connection

We adopted a multilevel feature fusion strategy to preserve detailed image information. Unlike U-Net3+, which directly upsamples and downsamples feature maps before merging, our full-scale skip connection integrates the previously described multi-attention gate to extract cross-level features more effectively. The encoders are denoted as E1–E5 and the decoders as D6–D9, with the corresponding formulas shown as Eqs. (1)–(4).

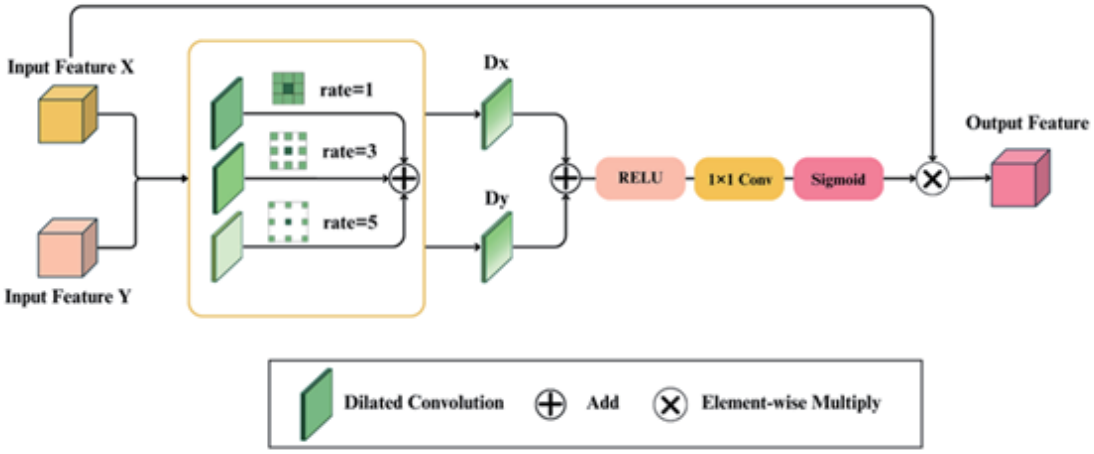

Fig. 3. (Color online) Multi-attention gate architecture.

$$D6 = H\left[MA\left(d(E1)^8, U(E5)^2\right) + MA\left(d(E2)^4, U(E5)^2\right) + MA\left(d(E3)^2, U(E5)^2\right) + MA\left(E4, U(E5)^2\right) + U(E5)^2 \right], \tag{1}$$

$$D7 = H\left[MA\left(d(E1)^4, U(D6)^2\right) + MA\left(d(E2)^2, U(D6)^2\right) + MA\left(E3, U(D6)^2\right) + MA\left(U(E4)^2, U(D6)^2\right) + MA\left(U(E5)^4, U(D6)^2\right) \right], \tag{2}$$

$$D8 = H\left[MA\left(d(E1)^2, U(D7)^2\right) + MA\left(E2, U(D7)^2\right) + MA\left(U(E3)^2, U(D7)^2\right) + MA\left(U(E4)^4, U(D7)^2\right) + MA\left(U(E5)^8, U(D7)^2\right) \right], \tag{3}$$

$$D9 = H\left[MA\left(E1, U(D8)^2\right) + MA\left(U(E2)^2, U(D8)^2\right) + MA\left(U(E3)^4, U(D8)^2\right) + MA\left(U(E4)^8, U(D8)^2\right) + MA\left(U(E5)^{16}, U(D8)^2\right) \right], \tag{4}$$

Here, $H(\cdot)$ represents the multi-res convolution block, and $d(\cdot)$ and $U(\cdot)$ denote downsampling and upsampling operations, respectively, with superscripts indicating scaling factors. $MA[x, y]$ denotes the inputs x and y of the multi-attention gate, and $\oplus$ indicates concatenation.

(1) Feature maps from each level are resized to match E4 via upsampling or downsampling. For example, E1 undergoes 8 × 8 max pooling and E2 undergoes 4 × 4 pooling.

(2) Feature maps from E1–E4 are paired with the upsampled E5 and input into the multi-attention gate to integrate high- and low-level semantic features.

(3) The five outputs are concatenated and passed through the multi-res block for convolution to extract features and restore channel depth. The same process is repeated for D7–D9.

2.4 Dual GTSA-CN

The Dual GTSA-CN architecture is designed to mimic the diagnostic logic employed by experienced orthopedic surgeons. In clinical practice, a surgeon typically identifies the presence and severity of GTS as a primary indicator before evaluating the overall joint condition for potential RCTs. To formalize this clinical intuition, our dual-branch framework employs a cross-modal feature guidance strategy. By utilizing the GTS mask generated in the first stage as a spatial prior, the GTSA module constrains the classification network's receptive field to focus on the most pathologically relevant areas. This design provides a stronger theoretical foundation than simple feature concatenation, as it explicitly models the causal relationship between specific localized bone changes and the global severity of the ligament tear.

In the classification stage, a dual-branch deep learning network (Fig. 4) is adopted, using both the X-ray image and the GTS mask generated in the segmentation stage as inputs. The network further classifies RCT cases into partial- and full-thickness tears.

2.5 Greater tuberosity sclerosis detection (GTSD) block

For the RCT region in X-ray images, a GTSD block is designed to detect the GTS region and assign weights, enabling the model to focus on the acromion and greater tuberosity. The GTSD block consists of two modules: GTSA and channel attention (CA).

(1) The X-ray input undergoes dilated and two standard convolutions, and the GTSA module generates an attention coefficient map.
(2) This coefficient map is multiplied by the 3 × 3 convolution output, integrating GTS mask information. To retain original image features, a residual connection adds the X-ray input to the 3 × 3 convolution output.
(3) The CA module performs channel selection, filtering useful features for classification. Max pooling reduces spatial size before passing features to the next layer.

2.6 Channel attention (CA) module

The CA module adaptively recalibrates the importance of different feature channels within the dual-branch sensing architecture. For a given feature map X, the module first performs

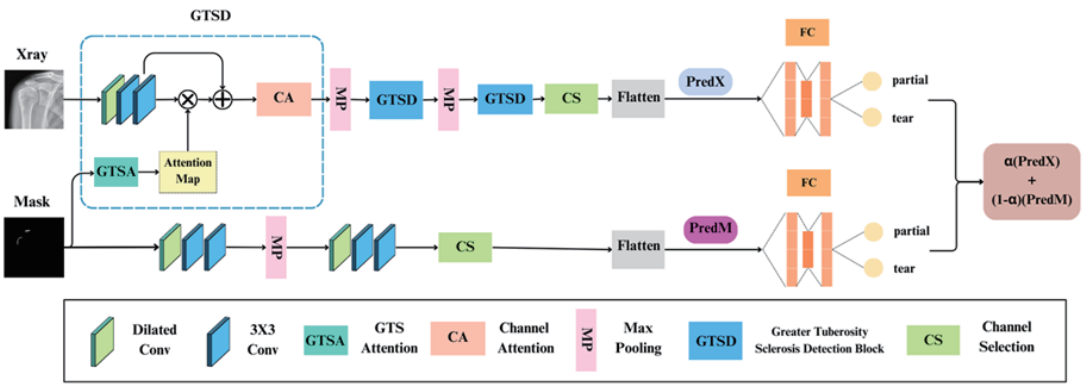

Fig. 4. (Color online) Dual GTSA-CN architecture.

global average pooling (GAP) to aggregate spatial information into a channel descriptor. This is followed by two 1 × 1 convolutional layers (bottleneck structure) with reduction ratio r=l6.

After spatial attention, a CA mechanism highlights important feature channels by considering global channel information, enhancing the model's ability to focus on globally relevant details. This produces a weight vector reflecting each channel's importance.

(1) The input feature map ($H \times W \times C$) undergoes 3 × 3 and 5 × 5 convolutions, and the resulting features are summed.
(2) The summed map passes through three adaptive average pooling layers of sizes 1 × 1 × C, 2 × 2 × C, and 4 × 4 × C, capturing global and multiscale local information.
(3) These outputs are flattened, concatenated, and passed through two fully connected layers and a sigmoid function to produce a channel weight vector (1 × 1 × C).
(4) The weight vector multiplies the input feature map to generate the final output.

In the X-ray image branch, three GTSD blocks are used. In the first block, the GTSA module takes the GTS mask as input, while in subsequent blocks, it uses the output of their own 3 × 3 convolutions.

2.7 GTSA module

The GTSA module uses a spatial attention mechanism to extract GTS information from the mask. The detailed operations are described as follows

(1) The GTSA module accepts the GTS mask M (dimensions $H \times W \times 1$) and processes it through a series of dilated convolutions with dilation rates of 1, 3, and 5. This allows the module to capture multiscale spatial features of the sclerotic lesions.
(2) These multiscale feature maps are combined through element-wise summation and activated by a rectified linear unit (ReLU) to enhance nonlinear feature mapping and accelerate convergence.
(3) A final 1 × 1 convolution is applied to adjust the channel count, followed by a sigmoid activation function to generate a normalized spatial attention map A (where all values range between 0 and 1). This attention map is then multiplied by the feature maps from the X-ray branch to prioritize pathologically relevant regions.

2.8 Prediction fusion

The X-ray image and GTS mask branches each flattens their final output feature maps after the channel selection module and feeds them into fully connected layers for binary classification, producing predictions PredX and PredM. The final output is obtained by fusing the two using

$$Pred_{final} = \alpha \cdot PredX + (1-\alpha) \cdot PredM, \tag{5}$$

where α is a weighting factor balancing the contributions of both branches. This fusion effectively integrates information from the X-ray images and GTS mask. Experiments showed that $\alpha = 0.4$ yielded the best performance.

2.9 Implementation details and loss functions

To train the proposed RCTSC framework, distinct loss functions were employed for each stage. For the segmentation stage (FSMA-SN), we used a hybrid loss function L_{seg} to balance pixel-level accuracy and region overlap, defined as

$$L_{seg} = \lambda_1 \cdot L_{BCE} + \lambda_2 \cdot L_{Dice}, \tag{6}$$

where λ_1 and λ_2 are weighting coefficients used to balance the contribution of each loss component. L_{BCE} represents the standard binary cross-entropy loss for pixel-wise classification, while L_{Dice} is the Dice loss, which specifically addresses the class imbalance between the subtle GTS regions and the background in X-ray sensor images.

For the classification stage (Dual GTSA-CN), the binary cross-entropy loss (L_{cls}) was utilized to optimize the differentiation between tear severities on the basis of the extracted sensing features:

$$L_{cls} = -\frac{1}{N} \cdot \sum \left[y_i \cdot \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \cdot \log\left(1 - \hat{y}_i\right) \right] \tag{7}$$

where N denotes the total number of samples in the training batch, y_i represents the ground truth label of the i-th sample (e.g., partial- vs full-thickness tear), and $\hat{y}_i$ is the predicted probability generated by the network. All models were implemented using Keras with a TensorFlow backend and optimized using the Adam optimizer with an initial learning rate of 10^{-4}.

3. Results

This study was conducted in collaboration with China Medical University Hospital, Taiwan. From January 2019 to December 2022, patients who underwent both conventional shoulder radiography and MRI within 90 days for shoulder pain were enrolled. Those with prior shoulder surgery, tumors, or infectious lesions were excluded. After reviewing complete X-ray and MRI series, 233 anteroposterior shoulder X-rays from 219 patients were collected. Low-quality images were excluded, and data augmentation through rotation and horizontal flipping produced 850 images—680 for training and 170 for testing.

The study was approved by the institutional review board (IRB: CMUH102-REX2-062). All data were anonymized, and no direct patient contact occurred. The research complied with institutional and national ethical standards and the Declaration of Helsinki. Both X-ray and ground truth images were resized to 512 × 512 pixels. The segmentation model output was binarized to remove noise and yield a more accurate GTS mask. For RCT classification, cases were divided by MRI findings into Normal, Partial-tear, and Full-thickness tear groups, with the latter two forming the tear group. The ground truth of GTS regions was annotated by two orthopedic surgeons.

To ensure the reliability of the experimental results and mitigate the risk of overfitting, several rigorous protocols were followed. First, the data split was strictly performed at the patient level; images from the same patient were assigned either to the training set or the test set, but never to both. This prevents the model from "memorizing" specific patient anatomical features and ensures true generalization. Second, data augmentation was limited to clinically plausible transformations—specifically, rotations within ±15 degrees and horizontal flipping—to simulate variations in patient posture during X-ray examinations. Furthermore, to prevent the "near-optimal" performance from being a result of overfitting, we incorporated dropout (p=0.5) in the classification layers and early stopping with a patience of 20 epochs during training. The high accuracy in segmentation is interpreted as a result of the effective suppression of background noise by the multi-attention gate, while the Dice coefficient provides a more conservative and accurate measure of lesion localization.

In this study, we evaluated the performance of both the segmentation and classification networks using standard metrics derived from the confusion matrix, including accuracy, precision, recall, specificity, F1-score, intersection over union (IoU), and Dice coefficient. The corresponding mathematical definitions of these metrics are presented as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}, \tag{8}$$

$$Precision = \frac{TP}{TP + FP}, \tag{9}$$

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{11}$$

$$F1\text{-}score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN}, \tag{13}$$

$$Dice = \frac{2TP}{2TP + FP + FN}. \tag{14}$$

Among these metrics, TP and TN represent the numbers of correctly predicted positive and negative samples, respectively, while FP and FN correspond to incorrect positive and negative predictions. Accuracy is the most intuitive performance measure but may be misleading when the dataset is imbalanced. Precision indicates how reliable the model's positive predictions are—high precision means few false positives. Recall is a measure of the model's ability to identify all positive samples—high recall implies few false negatives. The F1-score provides a harmonic

balance between precision and recall, making it a valuable metric when the relative importance of the two is uncertain.

We compared the proposed segmentation network with several state-of-the-art deep learning models. The results of comparison on the same dataset are summarized in Tables 1 and 2. As shown in Table 1, FSMA-SN achieved superior segmentation performance, with Dice and IoU scores of 0.835 and 0.704, respectively. Two factors contribute to this improvement. First, the multi-attention gate enhances the model's ability to capture fine details by fusing multilevel features through spatial attention in a pairwise manner. Second, the enhanced full-scale skip connection aggregates information from all feature levels, effectively reducing information loss during convolution. Finally, binarization of the output further aligns the prediction with the ground truth mask. To further evaluate the effectiveness of the segmentation stage, Attention U-Net was included as an additional state-of-the-art baseline. As shown in Table 1, while Attention U-Net improves upon the standard U-Net by focusing on salient features, the proposed FSMA-SN achieves a higher Dice coefficient (0.835) and IoU (0.704). This demonstrates the advantage of combining multi-attention gates with full-scale skip connections for capturing the subtle boundaries of sclerotic lesions. Furthermore, we compared our model with Swin-Unet, a representative transformer-based architecture for medical segmentation. As shown in Table 1, our proposed FSMA-SN maintains a higher Dice coefficient, suggesting that the integration of multiscale attention is more suited to extracting localized pathological signs like GTS than pure self-attention mechanisms in data-constrained scenarios.

During the segmentation stage, the generated output mask provides an initial binary diagnosis indicating the presence or absence of RCT. A completely black mask represents a normal shoulder X-ray, whereas any visible segmented region indicates an RCT lesion. As presented in Table 2, the classification performance was quantitatively evaluated in terms of accuracy, precision, recall, specificity, and F1-score. The proposed model achieved the highest accuracy (0.941) and F1-score (0.955) among all compared methods. In particular, the recall

Table 1
Results of quantitative comparison of segmentation performance among different models.

Model	Dice	IoU	Accuracy	Precision	Recall	Specificity	F1-score
U-Net[23]	0.790	0.625	0.998	0.902	0.751	0.999	0.792
U-Net++[30]	0.766	0.605	0.998	0.904	0.755	0.999	0.795
Eff-UNet[31]	0.79	0.569	0.998	0.833	0.712	0.999	0.752
Attention U-Net	0.812	0.678	0.998	0.895	0.782	0.999	0.814
Swin-Unet (Transformer)	0.821	0.690	0.998	0.880	0.805	0.999	0.841
Proposed	0.835	0.704	0.998	0.873	0.835	0.999	0.844

Table 2
Results of quantitative evaluation of segmentation-based classification performance among different models.

Model	Accuracy	Precision	Recall	Specificity	F1-score
U-Net[23]	0.876	0.858	0.963	0.73	0.907
U-Net++[30]	0.876	0.836	1	0.667	0.910
Eff-UNet[31]	0.818	0.839	0.879	0.714	0.858
Proposed	0.941	0.915	1	0.841	0.955

value of 1.000 confirms that the model successfully identified all RCT cases without omission. However, while a perfect recall reflects strong sensitivity, it does not necessarily indicate overall model excellence, as an increased recall may come at the expense of a higher false-positive rate. Therefore, a complementary evaluation through precision is essential to ensure balanced and reliable model performance.

To further illustrate the segmentation performance of FSMA-SN, a pixel-level confusion matrix was constructed on the basis of the predicted masks of the 170 test images (Table 3). Each image was a size of 512 × 512 pixels, resulting in a total of 44,564,480 evaluated pixels. From the obtained TP, FP, FN, and TN counts, the corresponding precision (0.873), recall (0.835), specificity (0.9993), and accuracy (0.9983) were computed. The derived Dice coefficient and IoU were 0.854 and 0.745, respectively. These corpus-level metrics differ slightly from the per-image averages reported in Table 1 (Dice = 0.835, IoU = 0.704), reflecting the well-known distinction between image-level and pixel-level aggregation: image-level metrics weight each image equally, whereas corpus-level metrics reflect the global pixel distribution. Both sets of results consistently demonstrate the strong segmentation performance of FSMA-SN.

In the second stage of the proposed framework, RCT images are further classified into partial-thickness and full-thickness tears. The performance of the proposed Dual GTSA-CN model was compared with other classification methods, and the results are summarized in Table 4. The proposed model achieved the best performance across all evaluation metrics for the following reasons. First, a dual-branch architecture was adopted to integrate both X-ray images and GTS masks as input, enabling complementary feature extraction. Second, the incorporation of GTSD modules effectively guides the model to focus on the most relevant regions, while the channel selection module optimizes feature representation prior to the fully connected layers. Finally, the fusion prediction strategy further enhances classification accuracy by aggregating multibranch outputs. As shown in Table 4, the proposed Dual GTSA-CN consistently outperforms existing methods in terms of accuracy, precision, recall, specificity, and F1-score. To provide a more comprehensive comparison, we further incorporated ResNet50 and the Vision Transformer (ViT) as competitive baselines. While ViT demonstrates strong global feature extraction capabilities and ResNet50 remains a robust standard, our proposed Dual GTSA-CN

Table 3
Pixel-level confusion matrix of FSMA-SN segmentation results.

Actual/predicted	Positive (Pred = 1)	Negative (Pred = 0)
Positive (Actual = 1)	226898 (TP)	44836 (FN)
Negative (Actual = 0)	33008 (FP)	44259738 (TN)

Table 4
Results of comparison of classification performance among different models.

Model	Accuracy	Precision	Recall	Specificity	F1-score
Cho *et al.*[25]	0.729	0.75	0.679	0.778	0.713
Hashimoto *et al.*[28]	0.626	0.592	0.793	0.463	0.677
ResNet50	0.781	0.792	0.765	0.801	0.778
Vision Transformer	0.803	0.811	0.820	0.785	0.815
Proposed	0.86	0.839	0.887	0.833	0.862

achieves superior results across all metrics. This superiority underscores the advantage of using explicit mask-based spatial guidance and segmented GTS masks as spatial priors; this focuses the network on the most diagnostically relevant regions rather than relying solely on raw radiographic features or global self-attention.

Figure 5 presents representative examples for comparing the proposed method with other approaches. As illustrated, the FSMA-SN model demonstrates superior segmentation performance by producing more precise GTS delineations with sharper boundaries and minimal artifacts or fragmentations. Furthermore, for normal cases without RCT, the segmentation output remains completely black, indicating the absence of GTS detection. In contrast, other models occasionally generate small white artifacts or residual regions, suggesting less effective suppression of false activations.

To assess the contribution of each component in the proposed method, an ablation study was performed; the results are presented in Table 5. The term "w/o FSC + MA" denotes the model without the full-scale skip connection (FSC) or the multi-attention gate (MA). The results demonstrate that the complete model achieves the best overall performance across all metrics. Both the FSC and MA modules substantially improve the segmentation precision and classification accuracy.

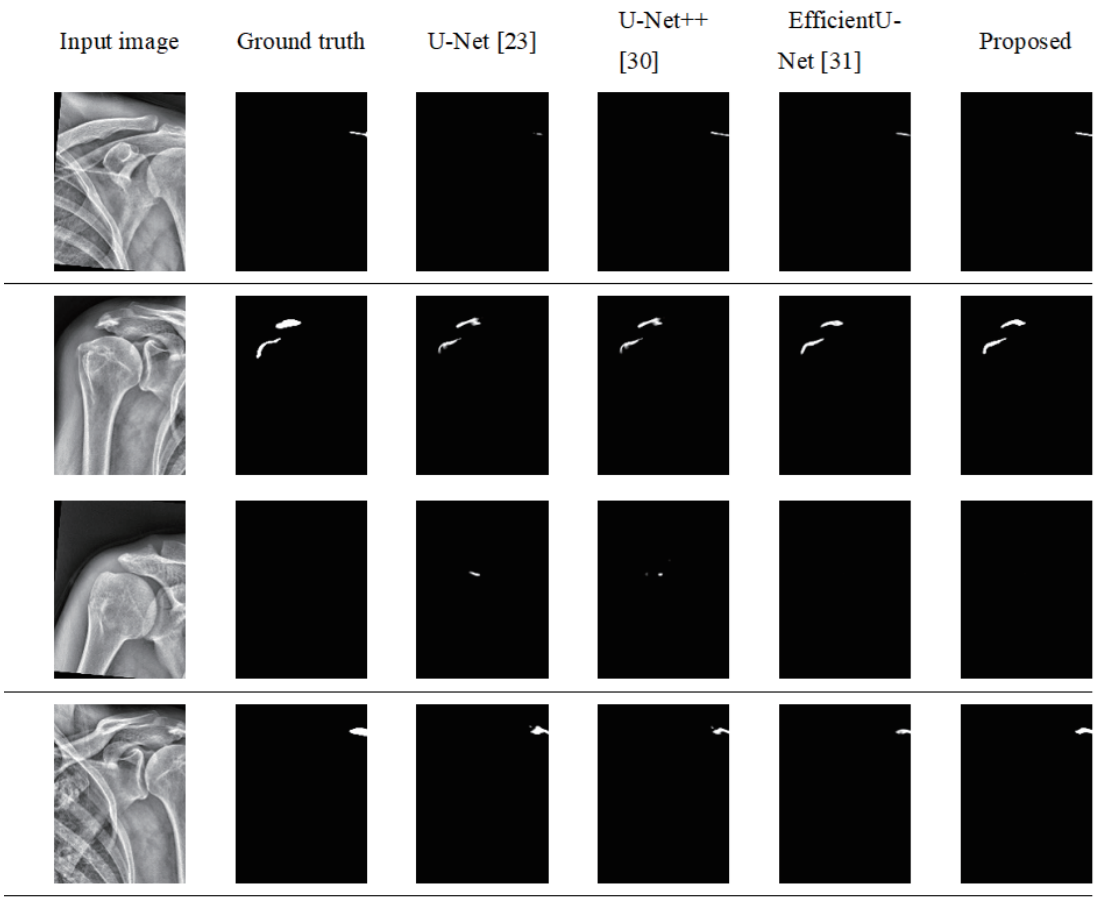

Fig. 5. Visual segmentation results obtained by the proposed FSMA-SN method and other approaches.

Table 5
Ablation study comparative data on the segmentation model based on Dice, IoU, and the confusion matrix.

Model	Dice	IoU	Accuracy	Precision	Recall	Specificity	F1-score
w/o FSC + MA	0.807	0.652	0.998	0.892	0.786	0.999	0.823
Proposed	0.835	0.704	0.998	0.873	0.835	0.999	0.844

To investigate the contribution of each component, ablation experiments were conducted on all modules of the Dual GTSA-CN, and their impact on classification performance is presented in Table 6. The abbreviation "w/o" indicates "without". Accordingly, "w/o GTSA, with CA" means the GTSA module was removed while the CA module remained, and "w/o CA, with GTSA" denotes the opposite configuration. Meanwhile, "w/o CS, with GAP" refers to replacing the channel selection module with the conventional GAP operation. The results reveal that both the GTSA and CA modules contribute to improved accuracy. Furthermore, replacing GAP with the channel selection module notably enhanced the performance, raising accuracy from 0.79 to above 0.8.

4. Discussion

In this study, we proposed a two-stage deep learning framework for the diagnosis of RCT using shoulder X-ray images, which comprises a segmentation stage and a classification stage. In the first stage, referring to the U-Net3+ architecture, we developed an enhanced segmentation network, FSMA-SN, which integrates a full-scale skip connection and a multi-attention mechanism. These improvements enable the model to effectively capture multiscale contextual information and enhance both segmentation accuracy and generalization performance. In the second stage, we designed a dual-branch classification model, Dual GTSA-CN, that takes both X-ray images and their corresponding GTS masks as inputs. This design compensates for the information insufficiency inherent in relying solely on X-ray images. The incorporation of the GTSA mechanism enhances feature extraction in the acromion and greater tuberosity regions, thereby improving classification accuracy. Moreover, the model employs an adaptive fusion strategy that balances predictions from both branches to achieve more reliable results.

Furthermore, to address the limitations of traditional GAP, which may lead to information loss, a channel selection module was introduced to dynamically recalibrate channel importance. This design preserves informative features while maintaining computational efficiency, enabling adaptability across different hardware environments. Ablation studies confirmed that each proposed component contributes positively to overall performance. Comparative evaluations with established deep learning models further demonstrate that the proposed framework achieves superior segmentation precision and classification accuracy. In addition, the proposed two-stage framework provides inherent interpretability by explicitly localizing GTS regions, which guide the subsequent classification process. From a clinical perspective, the system has the potential to assist physicians in detecting RCT more efficiently, thereby reducing reliance on MRI, shortening diagnostic waiting times, and lowering healthcare costs.

Despite these promising results, several limitations should be acknowledged. First, the dataset is relatively small and derived from a single institution, which may affect the generalizability of the model. Nevertheless, the proposed two-stage framework mitigates this

Table 6
Ablation study comparative data on classification models.

Models	Accuracy	Precision	Recall	Specificity	F1 score
w/o GTSA with CA	0.841	0.833	0.849	0.833	0.841
w/o CA with GTSA	0.813	0.811	0.811	0.815	0.811
w/o CS with GAP	0.794	0.782	0.811	0.778	0.796
Proposed	0.860	0.839	0.887	0.833	0.862

issue to some extent. The integration of segmentation and classification serves as an implicit regularization mechanism, guiding the model to focus on anatomically meaningful regions (i.e., GTS) and reducing the risk of overfitting under limited data conditions. Future work will include validation on larger, multi-institutional datasets to further assess robustness across diverse populations and imaging protocols.

Second, the current model primarily relies on GTS as the key radiographic indicator for RCT. However, additional imaging features may also contribute to diagnosis. Incorporating a broader range of radiographic characteristics can further enhance diagnostic performance. In addition, future studies may incorporate external validation and cross-validation strategies to further strengthen the reliability of the proposed framework.

5. Conclusions

In conclusion, in this study, we addressed the clinical challenge of achieving an early, cost-effective diagnosis of RCTs by maximizing the information extracted from conventional X-ray sensors. To overcome the limitations of manual interpretation and the high cost of MRI, we proposed a novel two-stage deep learning framework designed to augment radiographic sensing systems. The proposed method first employs FSMA-SN to accurately delineate GTS from X-ray sensor data, overcoming the difficulty of identifying subtle bone density variations. Subsequently, Dual GTSA-CN utilizes these segmented masks as spatial priors to guide the diagnosis, effectively mimicking the clinical workflow of orthopedic surgeons.

Experimental results demonstrate that this framework significantly outperforms existing state-of-the-art methods in both segmentation and classification tasks. In future work, we plan to focus on three key directions: (1) conducting extensive validations on large-scale, multicenter datasets to further ensure the model's generalization; (2) expanding the framework to incorporate other indirect radiographic signs captured by X-ray sensors, such as acromiohumeral interval narrowing or cystic changes; and (3) developing a clinical decision-support interface to evaluate the model's real-world utility in primary orthopedic care.

References

1 J. Lucas, P. Van Doorn, E. Hegedus, J. Lewis, and D. Van Der Windt: BMC Musculoskelet. Disord. **23** (2022) 1073. https://doi.org/10.1186/s12891-022-05973-8
2 C. Mitchell, A. Adebajo, E. Hay, and A. Carr: BMJ Open **331** (2005) 1124. https://doi.org/10.1136/bmj.331.7525.1124
3 S. N. Sambandam, V. Khanna, A. Gul, and V. Mounasamy: World J. Orthop. **6** (2015) 902. https://doi.org/10.5312/wjo.v6.i11.902

4 N. D. Clement, Y. X. Nie, and J. M. McBirnie: Sports Med. Arthrosc. Rehabil. Ther. Technol. **4** (2012) 48. https://doi.org/10.1186/1758-2555-4-48
5 V. Pandey and W. J. Willems: Asia-Pac. J. Sports Med. Arthrosc. Rehabil. Technol. **2** (2015) 1. https://doi.org/10.1016/j.asmart.2014.11.003
6 H. C. Chuang, C. K. Hong, K. L. Hsu, F. C. Kuan, C. H. Chiang, Y. Chen, and W. R. Su: JSES Int. **5** (2021) 77. https://doi.org/10.1016/j.jseint.2020.09.015
7 T. M. Ghandour, S. A. Elghazaly, and A. M. Ghandour: Acta Orthop. Belg. **83** (2017) 416.
8 M. Ghandour, S. El Ghazaly, and T. El Ghandour: Egypt. J. Radiol. Nucl. Med. **48** (2017) 425. https://doi.org/10.1016/j.ejrnm.2017.02.005
9 R. W. Westermann, C. Schick, C. M. Graves, K. R. Duchman, and S. L. Weinstein: Clin. Orthop. Relat. Res. **475** (2017) 580. https://doi.org/10.1007/s11999-016-5181-9
10 P. Sanguanjit, A. Apivatgaroon, P. Boonsun, S. Srimongkolpitak, and B. Chernchujit: Sci. Rep. **12** (2022) 9404. https://doi.org/10.1038/s41598-022-13632-0
11 M. Xu, Z. Li, Y. Zhou, B. Ji, S. Tian, and G. Chen: BMC Musculoskelet. Disord. **21** (2020) 106. https://doi.org/10.1186/s12891-020-3109-8
12 F. Suluova, U. Kanatli, B. Y. Ozturk, E. Esen, and S. Bolukbasi: Eur. J. Orthop. Surg. Traumatol. **24** (2014) 733. https://doi.org/10.1007/s00590-013-1247-5
13 J. Y. Gwark, T. S. Park, and H. B. Park: J. Orthop. Surg. **27** (2019). https://doi.org/10.1177/2309499019825762
14 Y. W. Pan: Shoulder Elbow **3** (2011) 205. https://doi.org/10.1111/j.1758-5740.2011.00143.x
15 D. Goutallier, J. M. Postel, J. Bernageau, L. Lavau, and M. C. Voisin: Clin. Orthop. Relat. Res. **304** (1994) 78.
16 D. A. Lansdown, S. Lee, C. Sam, R. Krug, B. T. Feeley, and C. B. Ma: Orthop. J. Sports Med. **5** (2017) 2325967117718537. https://doi.org/10.1177/2325967117718537
17 K.-C. Lee, Y. Cho, K.-S. Ahn, H.-J. Park, Y.-S. Kang, S. Lee, D. Kim, and C. H. Kang: Diagnostics **13** (2023) 3254. https://doi.org/10.3390/diagnostics13203254
18 E. Shim, J. Y. Kim, J. P. Yoon, S.-Y. Ki, T. Lho, Y. Kim, and S. W. Chung: Sci. Rep. **10** (2020) 15632. https://doi.org/10.1038/s41598-020-72357-0
19 D. Guo, X. Liu, D. Wang, X. Tang, and Y. Qin: J. Orthop. Surg. Res. **18** (2023) 426. https://doi.org/10.1186/s13018-023-03909-z
20 J. Cui, X. Xia, J. Wang, X. Li, M. Huang, S. Miao, D. Hao, and J. Li: Acad. Radiol. **31** (2024) 994. https://doi.org/10.1016/j.acra.2023.09.012
21 S. H. Lee, J. Lee, K. S. Oh, J. P. Yoon, A. Seo, Y. J. Jeong, and S. W. Chung: PLoS ONE **8** (2023) e0284111. https://doi.org/10.1371/journal.pone.0284111
22 M. A. Esfandiari, M. F. Tafti, N. J. Dabanloo, and F. Yousefirizi: Heliyon **9** (2023) e15804. https://doi.org/10.1016/j.heliyon.2023.e15804
23 O. Ronneberger, P. Fischer, and T. Brox: Proc. Int. Conf. Medical Image Computing and Computer-assisted Intervention – MICCAI 2015 (Springer, 2015) 9351. https://doi.org/10.1007/978-3-319-24574-4_28
24 H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. W. Chen, and J. Wu: Proc. ICASSP 2020 - 2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing (IEEE, 2020) 1055. https://doi.org/10.1109/ICASSP40776.2020.9053405
25 Y. Cho, A Jalics, D. Lv, M. Gilbert, K. Dickon, D. Chen, T. Nguyen, H. Joines, B. Kakos, C. Chen, and S. Lemos: Proc. 2021 10th Int. Conf. Computing and Pattern Recognition (ACM, 2021) 237. https://doi.org/10.1145/3497623.3497661
26 Y. Kim, D. Choi, K. J. Lee, Y. Kang, J. M. Ahn, E. Lee., J. W. Lee, and H. S. Kang: Eur. Radiol. **30** (2020) 2843. https://doi.org/10.1007/s00330-019-06639-1
27 S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon: Proc. European Conf. Computer Vision (ECCV 2018) (Springer, 2018) 3. https://doi.org/10.1007/978-3-030-01234-2_1
28 E. Hashimoto, S. Maki, N. Ochiai, S. Ise, K. Inagaki, Y. Hiraoka, F. Hattori, and S. Ohtori: J. Shoulder Elbow Surg. **33** (2024) 1733. https://doi.org/10.1016/j.jse.2023.12.009
29 M. Tan and Q. Le: Proc. 36th Int. Conf. Machine Learning (ICML 2019) (PMLR, 2019) 6105.
30 Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang: Proc. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA 2018, ML-CDS 2018) (Springer, 2018) 3. https://doi.org/10.1007/978-3-030-00889-5_1
31 B. Baheti, S. Innani, S. Gajre, and S. Talbar: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW) (IEEE, 2020) 1473. https://doi.org/10.1109/CVPRW50498.2020.00187