

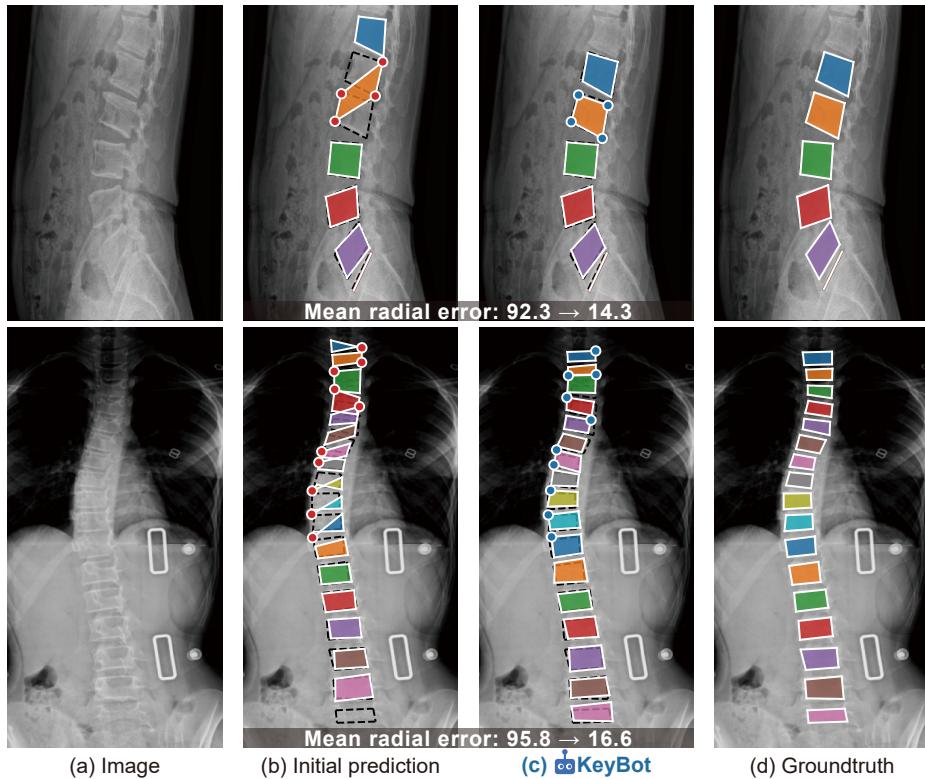
# Bones Can't Be Triangles: Accurate and Efficient Vertebrae Keypoint Estimation through Collaborative Error Revision

Jinhee Kim<sup>\*1</sup>, Taesung Kim<sup>\*1</sup>, and Jaegul Choo<sup>1</sup>

Korea Advanced Institute of Science and Technology

{seharanul17, zkm1989, jchoo}@kaist.ac.kr

\* indicates equal contributions.



**Fig. 1:** Vertebrae keypoint refinement results on BUU-LA [11] and AASCE [24]. Initial predictions exhibit significant errors in vertebra shapes due to difficulty identifying individual vertebrae. KeyBot effectively reduces these errors without user input.

**Abstract.** Recent advances in interactive keypoint estimation methods have enhanced accuracy while minimizing user intervention. However, these methods require user input for error correction, which can be costly in vertebrae keypoint estimation where inaccurate keypoints are densely clustered or overlap. We introduce a novel approach, KeyBot, specifically designed to identify and correct significant and typical errors in existing models, akin to user revision. By characterizing typical error types

and using simulated errors for training, KeyBot effectively corrects these errors and significantly reduces user workload. Comprehensive quantitative and qualitative evaluations on three public datasets confirm that KeyBot significantly outperforms existing methods, achieving state-of-the-art performance in interactive vertebrae keypoint estimation. The source code and demo video are available on our [project page](#).

**Keywords:** Interactive vertebrae keypoint estimation · X-ray image analysis · Automatic error correction · User interaction

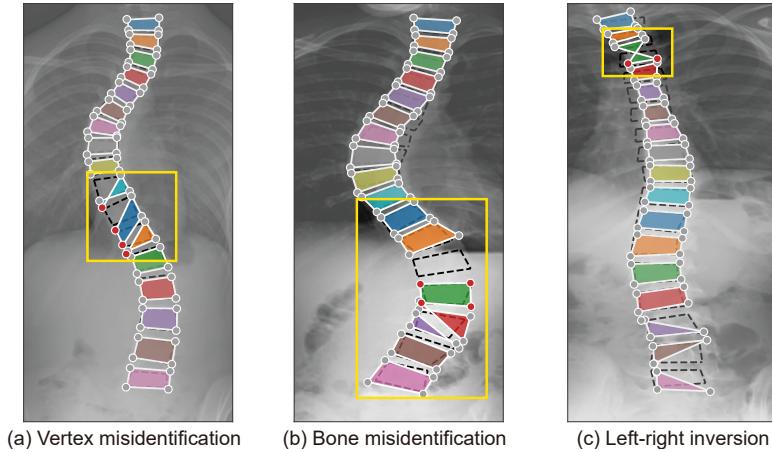
## 1 Introduction

Accurate vertebrae keypoint estimation from X-ray images is crucial for effective medical diagnosis and treatment planning [20, 23], with errors having significant impacts on clinical decisions, such as in spinal surgeries and pain management [6, 13, 15]. This task presents a distinct set of challenges due to the complex anatomy of the spine [17, 27]. The similar and repetitive shapes of vertebrae complicate distinguishing between individual vertebrae and accurately estimating their keypoints, as shown in Fig. 1. This requires a deep understanding of sequential context and anatomy, unlike more distinct keypoints found in human pose [5, 9, 14, 19] or facial keypoint [1, 3] estimation tasks.

Such challenges often result in significant errors during automated vertebrae keypoint estimation, including vertex misidentification, bone misidentification, and left-right inversion, as illustrated in Fig. 2. Manual revision of these errors is time-consuming and labor-intensive, especially as the number of vertebrae increases. Missing even a single vertebra can lead to extensive adjustments, thereby significantly increasing the time and effort required for accurate annotation. To mitigate these, interactive keypoint estimation approaches have been developed, significantly improving performance with minimal user input [10, 26]. These methods aim to automatically refine inaccurate keypoints with minimal user feedback, using the feedback as a hint to adjust the remaining errors, as shown in Fig. 3(a). However, they heavily rely on users to identify and correct errors, which can be time-consuming and labor-intensive, especially when erroneous keypoints are densely clustered or overlapped.

In response, we introduce a novel method named **KeyBot**, designed to autonomously detect and correct specific errors in vertebrae keypoint estimation before user evaluation, building upon existing interactive keypoint estimation models, as illustrated in Fig. 3(b). Initially, the interaction model predicts keypoints based on the input image. During refinement, KeyBot autonomously pre-evaluates these predictions, identifies specific errors, and provides corrective feedback similar to user input. In contrast to previous models requiring at least one user input, KeyBot autonomously refines model predictions before user input.

We generate synthetic data characterizing typical error types, vertex misidentification, bone misidentification, and left-right inversion, and train KeyBot using this data to focus specifically on predefined error types. Independently trained from the interaction model, KeyBot effectively identifies errors that might be



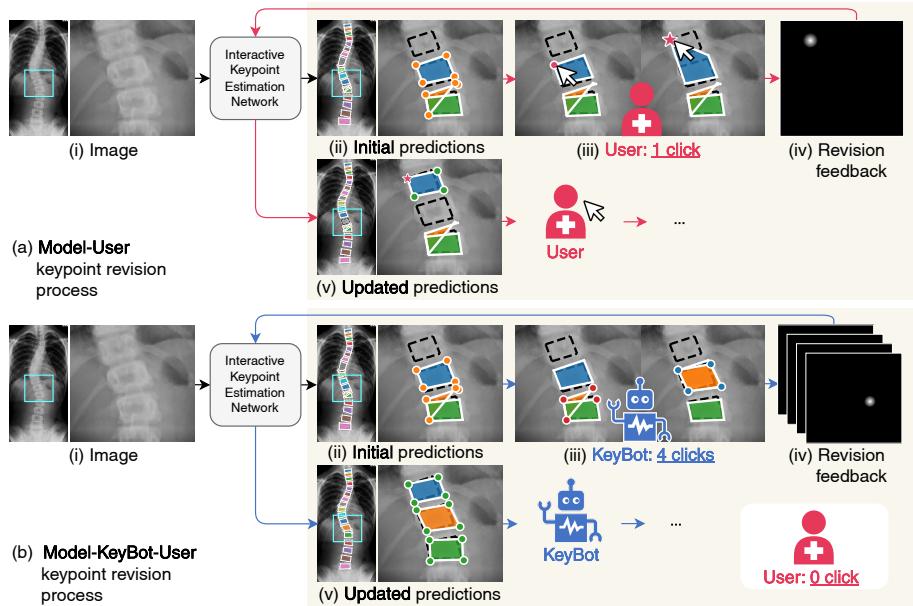
**Fig. 2:** Error types in vertebrae keypoint estimation, illustrated on the AASCE dataset.

overlooked by the interaction model, significantly reducing the user's revision efforts. Rather than correcting errors from scratch, users can focus on verifying and fine-tuning results, with KeyBot having already eliminated major errors.

KeyBot delivers quick, cost-effective, and iterative feedback without user input, which is crucial for complex cases requiring multiple refinements. This automated feedback is typically much faster than user review, significantly enhancing workflow efficiency. The collaborative feedback loop between the user, KeyBot, and the interaction model enriches the process by combining expert medical insights with KeyBot's focus on typical error types. This synergy addresses a broad range of potential errors and enhancements, leading to an overall improvement in keypoint estimation accuracy.

Through extensive experiments on three public datasets [11, 24], KeyBot demonstrates significant improvements in vertebrae keypoint estimation accuracy. The model reduces the mean radial error (MRE) by 19% and decreases the number of user clicks (NoC) required for target performance by 17% on the AASCE dataset, achieving state-of-the-art performance. These results highlight the significance of KeyBot in advancing vertebrae keypoint estimation, offering an advanced, user-centric solution to this unique and complex challenge.

In summary, our contributions are: (1) We introduce a novel method, KeyBot, designed to provide automated corrections for specific errors in vertebrae keypoint estimation. This enhances existing interactive keypoint estimation frameworks and improves overall keypoint estimation accuracy and efficiency. (2) We train KeyBot on synthetic data representing common error types in vertebrae, enabling it to effectively identify and address these errors. (3) KeyBot significantly reduces the need for user intervention during the keypoint annotation process. (4) Extensive evaluation across three public datasets confirms the efficacy of our proposed method in improving keypoint estimation accuracy and significantly reducing user intervention.



**Fig. 3:** Overview of (a) the interactive keypoint estimation and (b) the KeyBot framework. The interaction model generates initial keypoint predictions from an image, which are then revised by users, leading to updated results. This approach requires user input for error correction. In contrast, KeyBot offers a cost-efficient feedback mechanism via automated, rapid, and iterative refinement without user input. It independently identifies and corrects major errors, allowing users to focus on the final adjustments.

## 2 Related work

**Automated error refinement.** Despite the advancements in deep learning, errors in model predictions remain a formidable challenge. To mitigate these inaccuracies, a specialized area of research dedicated to refining prediction results has emerged, leading to the development of advanced methods for identifying and correcting errors in model outputs [4, 14, 16, 25]. For example, Qu et al. [16] introduced a method leveraging attention maps to detect label errors in CT data, markedly improving annotation efficiency and facilitating the collection of extensive datasets. Similarly, efforts to refine keypoint estimation have been noteworthy. For instance, Ronchi et al. [18] analyzed human pose estimation results to categorize three main types of errors, while Moon et al. [14] developed a method that specifically addresses and corrects these categorized errors, achieving state-of-the-art human pose estimation performances.

Despite these advancements, estimating vertebrae keypoints from X-ray images introduces additional challenges due to the complex sequential structure of the spine. Error patterns, such as the mislocalization of an entire vertebra, often require extensive user inputs for correction. In response, we propose KeyBot,

a method tailored for revising errors specific to vertebrae keypoint estimation, effectively addressing the unique challenges of this task.

**User interaction.** To maximize the accuracy of predictions while minimizing the need for user input, the development of interaction-based models is paramount. Interactive segmentation is a particularly active area within this domain. Backpropagating refinement schemes such as BRS [8] and f-BRS [21] have improved interactive image segmentation by incorporating user hints through backpropagation. Similarly, RITM [22] has shown superior performance by integrating an iterative refinement process into its training process. In the field of keypoint estimation, interactive keypoint estimation methods have been researched to reduce user input through interactive refinement frameworks. Kim et al. [10] pioneered an interactive framework for keypoint estimation, significantly improving the accuracy and efficiency of the keypoint annotation process in medical datasets. ClickPose [26] developed an interactive keypoint estimation model for multi-person 2D keypoints, significantly reducing annotation costs.

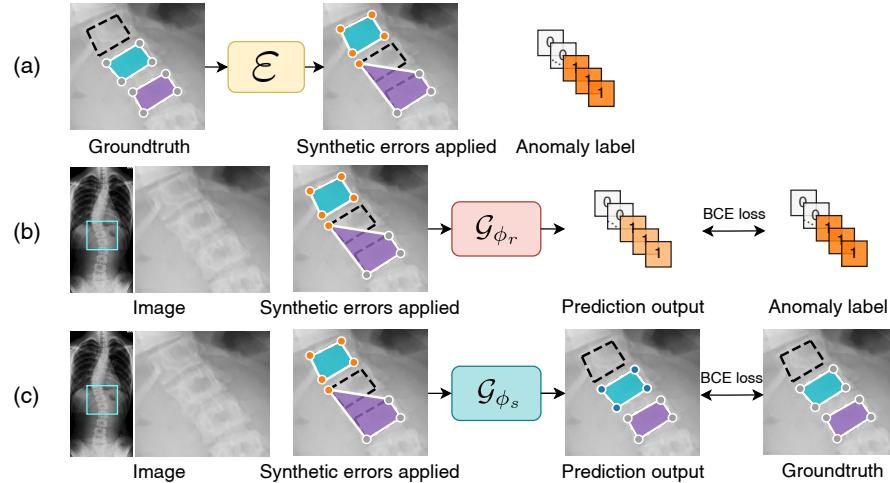
Recently, Liu et al. [12] introduced a method that generates pseudo clicks to mimic human interactions for refining segmentation predictions autonomously, thereby minimizing the required number of user clicks. However, their error decoder is trained end-to-end and can thus be affected by inherent biases or limitations of the segmentation module. To address this challenge, our KeyBot is trained independently to correct simulated errors, which allows KeyBot to provide corrective feedback free from any potential biases present in the keypoint estimation model. Unlike traditional interactive keypoint estimation models that rely on users to identify and correct errors, KeyBot autonomously generates corrective feedback. This significant innovation not only enhances the keypoint estimation accuracy but also remarkably reduces the need for user interaction, thereby streamlining the error correction process.

### 3 KeyBot

We introduce KeyBot, a novel method designed to enhance vertebrae keypoint estimation by identifying and correcting errors prior to user intervention. We employ an existing interactive model that can incorporate user feedback for interactive modifications. Given model predictions, KeyBot identifies erroneous keypoints and generates their pseudo-corrections, which are then fed to the interaction model for updates. KeyBot consists of two components: The detector discerns whether each input keypoint is accurate, and the corrector refines input keypoints accurately. Both components are trained with synthetic errors to focus specifically on the target error types.

#### 3.1 Overview of the propose approach

**Interactive keypoint estimation framework.** Consider an interactive keypoint estimation model  $\mathcal{F}_\theta$  with parameters  $\theta$ , operating on an image  $\mathbf{x}$ . The model aims to precisely estimate  $K$  target keypoints through user-interactive



**Fig. 4:** Training overview of KeyBot. KeyBot consists of two main components, the detector and the corrector. (a) Synthetic errors are introduced to the groundtruth keypoints, creating inaccurate keypoints and corresponding anomaly labels. (b) The detector is trained to discern whether each input keypoint is accurate or not. (c) The corrector is trained to refine these inaccurate keypoints accurately.

refinement. In each iteration  $t$ , the model generates the  $t$ -th prediction,  $\mathbf{y}_t$ , using three inputs: the image  $\mathbf{x}$ , user revisions  $\mathbf{u}_t$ , and its previous predictions  $\mathbf{e}_t$ :

$$\mathbf{y}_t = \mathcal{F}_\theta(\mathbf{x}, \mathbf{u}_t, \mathbf{e}_t), \quad (1)$$

where  $\mathbf{u}_0 := 0$  and  $\mathbf{e}_0 := 0$ . Subsequently, a user reviews the results, and the adjustment made to a keypoint position is encoded as Gaussian heatmaps  $\mathbf{u}_t^i$ , effectively capturing and representing the refined locations:

$$\rho_{t+1}, \mathbf{u}_{t+1} = \Psi(\mathbf{x}, \mathbf{y}_t, \mathbf{u}_t), \quad (2)$$

where the set  $\rho_t$  represents keypoint indices revised by the user over  $t$  iterations. Next, we selectively gather the *known-to-be-false* predictions from  $\mathbf{y}$  to  $\mathbf{e}$  and reintroduce them as inputs in the next step to differentiate mispredictions [10]:

$$\mathbf{e}_{t+1}^i = \mathbf{y}_t^i \text{ for } i \in \rho_t. \quad (3)$$

Here,  $\mathbf{e}$  denotes predictions modified by the user. These updates on  $\mathbf{u}_t$  and  $\mathbf{e}_t$  are critical for refining the model predictions, leading to the updated keypoint predictions reflecting the user feedback:

$$\mathbf{y}_{t+1} = \mathcal{F}_\theta(\mathbf{x}, \mathbf{u}_{t+1}, \mathbf{e}_{t+1}). \quad (4)$$

This process repeats until sufficient accuracy is achieved or the predefined maximum user interaction  $T$  is reached.

**The KeyBot framework.** KeyBot offers an automated keypoint refinement step that precedes and complements the existing user-based refinement, effectively reducing the need for user interaction in the process. KeyBot consists of the detector  $\mathcal{G}_{\phi_r}$  and corrector  $\mathcal{G}_{\phi_s}$ , which together identifies and corrects inaccurate keypoints predicted by  $\mathcal{F}_\theta$ . Following each prediction iteration of  $\mathcal{F}_\theta$ , KeyBot evaluates  $\mathbf{y}_t$  and suggests necessary corrections, identifying erroneous keypoints  $\boldsymbol{\nu}_{t,n+1}$  and their pseudo corrections  $\mathbf{z}_{t,n+1}$ :

$$\boldsymbol{\nu}_{t,n+1} = \mathcal{G}_{\phi_r}(\mathbf{x}, \mathbf{y}_{t,n}) \text{ and } \mathbf{z}_{t,n+1} = \mathcal{G}_{\phi_s}(\mathbf{x}, \mathbf{y}_{t,n}), \quad (5)$$

with  $\mathbf{y}_{t,0} := \mathbf{y}_t$ . This step is similar to the user interaction step outlined in Eq. (2). Here, KeyBot can detect multiple inaccurate keypoints at once and simultaneously offer corrective feedback for each identified error. If no inaccuracies are detected by the detector, KeyBot concludes and  $\mathbf{y}_{t,n}$  becomes the final prediction. Revision information,  $\mathbf{c}_{t,n}$ , now includes both user ( $\mathbf{u}_{t,n}$ ) and model ( $\mathbf{z}_{t,n}$ ) corrections. Here, user modifications are prioritized as the definitive ground truth. Thus, keypoints already revised by the user are exempt from further updates by the pseudo-corrections from KeyBot.

$$\mathbf{c}_{t,n+1}^i = \mathbf{z}_{t,n+1}^i \text{ for } i \in \boldsymbol{\nu}_{t,n+1} \setminus \boldsymbol{\rho}_t. \quad (6)$$

Previous work [10, 22] relies solely on the prediction at the most recent step as the previous prediction, as described in Eq. (3). However, this approach can result in losing track of erroneous predictions once they are corrected in the model. Hence, our approach accumulates predictions identified as erroneous by either KeyBot or the user across all iterations. These false predictions retain the history of incorrect predictions and serve to inform the model about specific prediction errors to avoid, thereby guiding it toward more accurate future predictions:

$$\mathbf{e}_{t,n+1}^i = \mathbf{y}_{t,n}^i \text{ for } i \in \boldsymbol{\nu}_{t,n+1} \setminus \boldsymbol{\nu}_{t,n} \setminus \boldsymbol{\rho}_t. \quad (7)$$

Here,  $\mathbf{e}$  denotes predictions modified by either the user or KeyBot. Finally, these are then fed back into the backbone interaction network for a new, corrected prediction akin to user feedback:

$$\mathbf{y}_{t,n+1} = \mathcal{F}_\theta(\mathbf{x}, \mathbf{c}_{t,n+1}, \mathbf{e}_{t,n+1}). \quad (8)$$

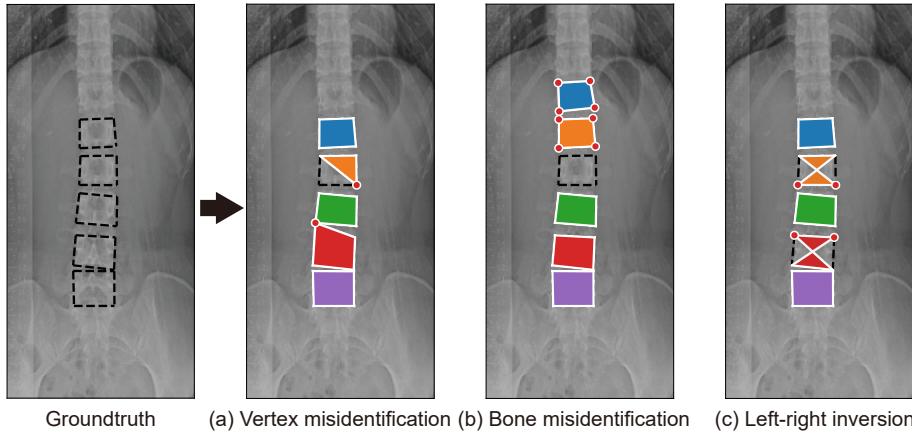
Each iteration of KeyBot contributes progressively towards enhancing the overall accuracy and reliability of the keypoint predictions. This cycle repeats until the detector detects no errors or the maximum KeyBot iteration count  $N$  is reached. Upon the KeyBot process concludes, say, after  $n+1$  iterations, the prediction  $\mathbf{y}_{t,n+1}$  is established as the final output for the  $t$ -th iteration of the interactive keypoint estimation process. For the subsequent  $(t+1)$ -th user-side phase, given new user feedback as Eq. (2),  $\mathbf{c}_{t,n+1}$  and  $\mathbf{e}_{t,n+1}$  are updated to  $\mathbf{c}_{t+1,0}$  and  $\mathbf{e}_{t+1,0}$ ,

$$\mathbf{c}_{t+1,0} = \mathbf{u}_{t+1} \text{ and } \mathbf{e}_{t+1,0}^i = \mathbf{y}_{t,n+1}^i \text{ for } i \in \boldsymbol{\rho}_{t+1}, \quad (9)$$

and the keypoints are revised accordingly:

$$\mathbf{y}_{t+1,0} = \mathcal{F}_\theta(\mathbf{x}, \mathbf{c}_{t+1,0}, \mathbf{e}_{t+1,0}). \quad (10)$$

The complete procedure is summarized in Algorithm 1 in Appendix B.



**Fig. 5:** Implementing realistic estimation inaccuracy simulations for training KeyBot.

### 3.2 Model design of KeyBot

KeyBot integrates the detector and corrector to identify potential errors and generates corrective feedback for their revision.

**Error detector.** The error detector specializes in identifying up to  $K$  erroneous keypoints, focusing on prevalent error types in vertebrae keypoint estimation. Our empirical analysis indicates that evaluating the entire bone structure at once is less effective for error detection. Therefore, the detector evaluates keypoints in smaller groups, examining  $k \leq K$  keypoints simultaneously. This approach simplifies the task and improves its performance. During training, we randomly select a set of consecutive groundtruth keypoints  $\mathbf{y}^*$  and either maintain their accuracy or introduce errors. The cropped image area around these keypoints, along with their corresponding Gaussian keypoint heatmaps, are input to the detector. It then predicts the likelihood of error presence in each keypoint.

Let  $\varepsilon$  denote the function that induces synthetic errors in groundtruth keypoints, and  $v$  represents the anomaly labels for the keypoints. The detector performs multi-label binary classification to differentiate between normal and abnormal keypoints using a sigmoid function in its output layer. It is supervised to minimize the Binary Cross-Entropy (BCE) loss between the anomaly labels  $v$  and the predictions:

$$\phi_r \leftarrow \phi_r - \eta_r \nabla_{\phi_r} \mathcal{L}_{bce}(v, \mathcal{G}_{\phi_r}(\mathbf{x}, \varepsilon(\mathbf{y}^*))), \quad (11)$$

with  $\eta_r$  representing the learning rate for the detector.

During inference, the detector iteratively detects errors across all keypoints, examining  $k$  keypoints at a time with a stride of  $s$ . To facilitate a comprehensive assessment of each keypoint in different contexts, we set the stride  $s$  to be smaller than  $k$ . This results in multiple predictions for a single keypoint, and if any prediction indicates an anomaly, the keypoint is flagged as erroneous.

**Error corrector.** The error corrector aims to generate pseudo-corrections for keypoints inaccurately identified by the detector, thus providing feedback to the model  $\mathcal{F}_\theta$ . During training, we randomly transform the groundtruth keypoints to erroneous ones using  $\varepsilon$ . Taking these keypoints and the image  $\mathbf{x}$  as inputs, the corrector is trained to revert these to their original, accurate state, focusing particularly on the typical error types implemented by  $\varepsilon$ . It generates  $K$ -channel heatmaps indicating reconstructed keypoint locations, employing a sigmoid function in its output. Its learning objective is the BCE loss between the ground truth keypoint heatmaps and the reconstructed outputs:

$$\phi_s \leftarrow \phi_s - \eta_s \nabla_{\phi_s} \mathcal{L}_{bce}(\mathbf{y}, \mathcal{G}_{\phi_s}(\mathbf{x}, \varepsilon(\mathbf{y}^*))), \quad (12)$$

where  $\eta_s$  is the learning rate for the corrector.

During inference, the corrector refines predictions from the backbone model  $\mathcal{F}_\theta$ , and the output corresponding to inaccurately assessed keypoints by the detector is fed to  $\mathcal{F}_\theta$  to serve as pseudo-corrections.

### 3.3 Training KeyBot via realistic error simulation

KeyBot aims to address significant and typical errors in vertebrae keypoint estimation. The key idea is to train KeyBot to identify and correct three error types commonly observed in vertebrae keypoint estimation: vertex misidentification, bone misidentification, and left-right inversion. During training, KeyBot is provided with both accurate keypoints and keypoints with synthetic errors. We randomly simulate the predefined error types, as shown in Fig. 5. Rather than relying on highly variable and not clearly defined real errors, we use characterized and consistent synthetic errors.

By employing this training approach, we expose KeyBot to a wide range of error scenarios, equipping it with the proficiency to detect and correct these errors effectively. As a result, KeyBot provides a targeted evaluation of the keypoints predicted by the interaction model, focusing on specific error types.

**Vertex misidentification (misvertex).** Misvertex is a type of error where a keypoint is correctly identified on a vertex of a vertebrae but is incorrectly attributed to the wrong vertex. For instance, a keypoint intended for the first vertex is inaccurately predicted at the third vertex. To simulate misvertex, we randomly displace a keypoint to its neighbors within a  $[-r, +r]$  index range for a predefined number  $r$ , excluding its original position. Keypoints moving beyond their index range wrap around cyclically.

**Bone Misidentification (misbone).** Misbone errors represent errors where the entire vertebra is incorrectly classified. Our simulation randomly shifts keypoints up or down by one vertebra. This is achieved by selecting start and end keypoint indices and moving all-encompassed keypoints to an adjacent vertebra.

**Left-right inversion (lr-inversion).** LR-inversion occurs when keypoint pairs are incorrectly positioned on the opposite lateral side. To simulate LR-inversion in our model, we randomly interchange the positions of keypoint pairs that delineate one side of each vertebra.

**Table 1:** Performance comparison of mean radial error in keypoint estimation across three datasets. UC denotes the count of user clicks provided to the model.

Method	AASCE					BUU-AP			BUU-LA		
	-	UC1	UC2	UC3	UC4	-	UC1	UC2	-	UC1	UC2
BRS [8]	45.65	35.88	29.93	25.40	21.97	51.22	30.91	36.17	40.20	43.19	35.07
f-BRS [21]	64.06	57.30	52.53	47.83	43.79	44.05	24.99	17.52	36.03	27.93	19.29
RITM [22]	56.03	37.96	31.03	26.76	23.79	36.43	25.76	17.96	23.27	17.79	9.89
Kim et al. [10]	51.58	30.60	25.78	21.60	19.08	42.31	23.26	16.46	23.43	14.29	10.29
Click-Pose [26]	54.65	46.50	44.08	41.73	40.04	32.72	29.30	26.20	33.70	21.62	17.38
KeyBot-i1	44.18	<b>25.93</b>	<b>20.66</b>	<b>18.03</b>	16.37	32.01	21.79	15.97	18.77	13.39	9.12
KeyBot-i2	42.52	27.03	22.95	18.72	<b>16.23</b>	31.88	<b>20.65</b>	<b>15.81</b>	19.11	13.47	9.03
KeyBot-i3	<b>41.70</b>	26.59	25.02	21.23	16.76	<b>31.87</b>	20.66	15.84	<b>18.74</b>	<b>13.36</b>	<b>8.97</b>

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We rigorously validate our method across three publicly available real-world datasets. The **AASCE** dataset [24] consists of 608 spinal anterior-posterior radiographs, each annotated with 68 keypoints, marking the vertices of 17 vertebrae. The BUU dataset [11] features two distinct subsets from 400 unique patients: **BUU-AP** with 400 anterior-posterior (AP) view radiographs, each annotated with 20 keypoints, and **BUU-LA** with 400 left lateral (LA) view radiographs, each annotated with 22 spinal keypoints. Images that exhibit critical keypoint annotation errors, such as those with incorrectly ordered keypoint indices similar to typical errors in keypoint estimation models, are excluded from the datasets. Comprehensive details about the dataset, including the data split and examples of the excluded images, are available in Appendix C.2.

**Metrics.** Following the evaluation protocol of previous work [10, 26], we evaluate the model performance using the mean radial error (MRE) and the number of clicks (NoC). MRE computes the mean Euclidean distance between the groundtruth and predicted keypoints. NoC<sub>a</sub>@b measures the average number of user clicks required to achieve a target MRE of b, given a maximum of a clicks.

**Baseline models.** We compare our method with state-of-the-art interactive keypoint estimation models, including the model proposed by Kim et al. [10] and Click-Pose [26]. Additionally, we use interactive image segmentation models such as BRS [8], f-BRS [21], and RITM [22] as baselines.

**Implementation details.** We utilize ResNet [7]-18 and DeepLab-V3 [2] as the backbone for the detector and corrector, respectively. We build on the pretrained model of Kim et al. [10], which is designed to estimate keypoints based on user feedback. All of the models are trained independently. We apply the misvertex error type to train the detector and all three error types for the corrector. KeyBot-i1, KeyBot-i2, and KeyBot-i3 refer to KeyBot with one, two, and three

**Table 2:** Comparison of required number of user clicks to reach target performance on the AASCE, BUU-AP, and BUU-LA datasets.

Method	AASCE				BUU-AP				BUU-LA			
	NoC <sub>10</sub>		NoC <sub>10</sub>		NoC <sub>5</sub>		NoC <sub>5</sub>		NoC <sub>5</sub>		NoC <sub>5</sub>	
	@20	@30	@40	@50	@15	@20	@25	@30	@15	@20	@25	@30
BRS [8]	2.53	1.97	1.51	1.23	2.80	1.51	0.92	0.61	3.89	2.26	1.34	0.86
f-BRS [21]	7.16	5.30	4.21	3.48	2.23	1.18	0.68	0.48	1.24	0.61	0.44	0.40
RITM [22]	3.20	2.37	1.87	1.38	1.05	0.58	0.42	0.30	0.57	0.24	0.17	0.14
Kim et al. [10]	2.10	1.54	1.23	1.03	0.94	0.54	0.37	0.32	0.38	0.15	0.14	0.12
Click-Pose [26]	4.15	3.56	3.20	2.97	1.20	0.89	0.76	0.71	1.45	0.70	0.56	0.45
KeyBot-i1	1.84	1.26	0.92	<b>0.53</b>	<b>0.87</b>	0.51	0.33	0.28	<b>0.34</b>	<b>0.10</b>	<b>0.06</b>	<b>0.05</b>
KeyBot-i2	1.78	<b>1.21</b>	<b>0.91</b>	0.61	<b>0.87</b>	<b>0.48</b>	<b>0.30</b>	<b>0.24</b>	0.35	0.11	<b>0.06</b>	<b>0.05</b>
KeyBot-i3	<b>1.74</b>	1.32	0.94	0.68	<b>0.87</b>	0.49	0.32	<b>0.24</b>	<b>0.34</b>	<b>0.10</b>	<b>0.06</b>	<b>0.05</b>

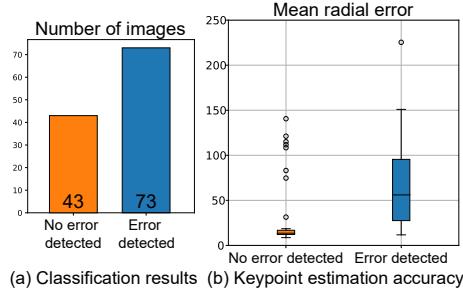
iterations, respectively. Unless explicitly stated, the experimental results are conducted without user interaction. The training and inference are conducted on a single NVIDIA GeForce RTX 3090 GPU with 24GB memory. More details on experimental settings, including training details, are provided in Appendix C.

## 4.2 Experimental results and analysis

**Comparison with state-of-the-art methods.** In this section, we present a comparative analysis of interactive keypoint estimation performance between KeyBot and established baseline models by increasing the maximum KeyBot iteration count  $N$  from one to three. As shown in Table 1, KeyBot achieves state-of-the-art performance in mean radial error (MRE) across the AASCE, BUU-AP, and BUU-LA datasets. KeyBot consistently outperforms baseline models both in initial predictions and during subsequent user interactions. Remarkably, KeyBot with a single user click (UC1) shows superior performance compared to BRS, f-BRS, RITM, and Click-Pose with two user clicks (UC2) on the AASCE dataset. Furthermore, the performance of KeyBot improves with additional iterations, demonstrating its ability to refine predictions through multiple correction cycles. These results highlight the effectiveness of KeyBot in enhancing keypoint estimation accuracy and reducing user intervention.

In terms of the number of clicks (NoC), as shown in Table 2, KeyBot exhibits superior performance compared to the baseline models across all three datasets. For instance, on the AASCE dataset, our proposed model nearly halves the NoC required to achieve a target MRE of 50 (NoC<sub>10</sub>@50) compared to that of the baseline models. This substantial reduction in NoC underscores KeyBot's proficiency in interactive frameworks, significantly lowering annotation costs.

**Analysis on the detector.** We investigate the effectiveness of the detector in identifying erroneous keypoints within images. In Fig. 6, we provide (a) the number of images where the detector identifies no errors versus those where errors are detected, as well as (b) the distribution of their actual mean radial error.



**Fig. 6:** Effectiveness of the detector in identifying errors on the AASCE dataset.

**Table 3:** Keypoint prediction errors on the AASCE dataset, comparing MRE for examples where the detector identifies errors versus where it does not.

Method	No error	Error
	detected	detected
Initial prediction	28.99	64.88
KeyBot-i1	-	53.12
KeyBot-i2	-	50.49
KeyBot-i3	-	49.18

Out of the total images analyzed, 43 images have no errors detected, while 73 images have errors identified. The images classified as “No error detected” exhibit significantly lower errors compared to those classified as “Error detected.” These findings indicate the precision and effectiveness of the detector in distinguishing between erroneous predictions and accurate ones.

We further examine error reduction through successive KeyBot iterations for images where errors are initially detected. As demonstrated in Table 3, KeyBot effectively enhances the accuracy of these error-detected images, significantly reducing errors in subsequent iterations. Notably, after three iterations of KeyBot, the MRE significantly decreases from 64.88 to 49.18. This reduction highlights KeyBot’s ability to iteratively identify inaccuracies and improve predictions.

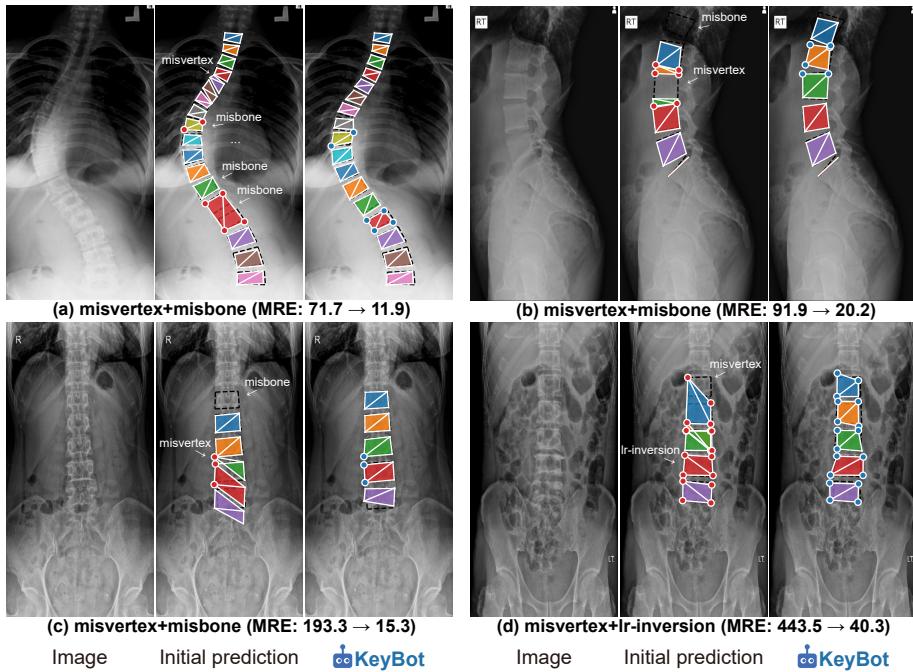
**Ablation study.** This section validates each component of KeyBot. First, we study the role of the detector under two scenarios: (i) employing corrector outputs directly as the final results and (ii) utilizing all corrector outputs as pseudo-corrections for the interaction model. As detailed in Table 4, both scenarios exhibit significant performance degradation compared to KeyBot, underscoring the importance of using the detector to selectively apply corrector outputs.

Additionally, utilizing only the interactive keypoint estimation network, without the support of KeyBot, results in a significant performance degradation compared to the refined predictions with KeyBot. When ablating the error types simulated during the training of the corrector, our findings show that training the corrector to recognize and rectify three distinct types of errors significantly improves its prediction accuracy compared to when it is trained on only one error type. This finding highlights the necessity of incorporating all three typical errors into the training process.

**Qualitative results.** In Fig. 7, we analyze the qualitative results of KeyBot with three iterations. The results demonstrate that KeyBot effectively identifies and reduces all three error types: vertex misidentification (Misvertex), bone misidentification (Misbone), and left-right inversion (LR-inversion). For these errors, KeyBot provides corrective feedback to the interactive model, enabling necessary adjustments. For instance, in Fig. 7(a), the initial prediction skips a bone, resulting in a Misbone error (red), and also exhibits a Misvertex error

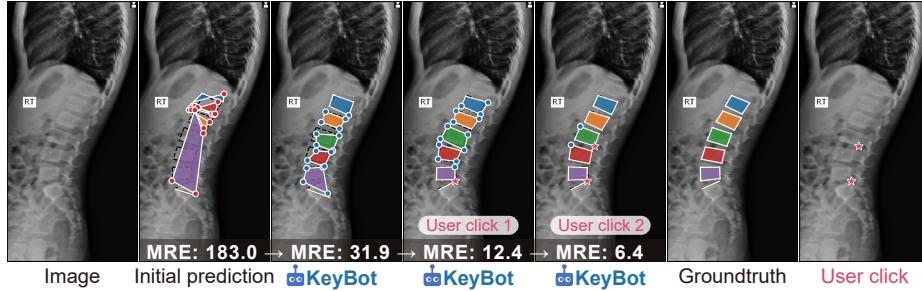
**Table 4:** Ablation study of our keypoint refinement method on the AASCE dataset.

Interaction model	KeyBot			Mean radial error -	Mean radial error			
	Detector	Corrector	Error type		UC1	UC2	UC3	UC4
✓	✓	✓	all	44.18	25.93	20.66	18.03	16.37
✗	✗	✓	all	50.99	39.17	32.98	30.36	28.57
✓	✗	✓	all	57.61	45.94	39.11	36.01	33.26
✓	✗	✗	-	51.58	30.60	25.57	21.07	18.55
✓	✓	✓	misvertex	52.67	31.14	25.92	21.65	18.52
✓	✓	✓	misbone	49.07	29.85	23.43	20.75	18.05
✓	✓	✓	lr-inversion	52.38	30.16	25.13	20.90	18.14



**Fig. 7:** Prediction results on the (a) AASCE, (b) BUU-LA, and (c,d) BUU-AP (bottom) datasets. KeyBot effectively reduces all three types of errors, even in challenging cases where error types are mixed. The keypoints classified as inaccurate by our detector are indicated with red dots, while their corresponding keypoints in the revised results are visualized as blue dots. A diagonal line connecting the upper-right vertex to the lower-left vertex is indicated in each vertebra. More results are in Appendix F.

(purple). KeyBot accurately refines these errors, significantly reducing the MRE from 71.7 to 11.9. The results highlight KeyBot’s capability to significantly enhance keypoint estimation accuracy without user clicks.



**Fig. 8:** Collaborative keypoint refinement process among the interaction model, KeyBot, and the user. Faced with substantial errors in initial predictions, KeyBot effectively removes all major errors, allowing users to easily identify and correct remaining inaccuracies with minimal effort. Here, refinement is achieved with just two user clicks.

Additionally, KeyBot excels in interpretability. It includes a separate detector that enables users to easily comprehend the keypoints identified as errors and the corrections made, leading to efficient evaluation of the results. In the given examples, identified errors are indicated by red dots, and the corrections are shown as blue dots, illustrating how the keypoints are adjusted. This transparency contributes to improving the efficiency of the overall annotation process. **Collaborative refinement process.** In Fig. 8, we showcase the qualitative results of the collaborative refinement process involving both KeyBot and the user. Our approach establishes a feedback loop among the interaction model, KeyBot, and the user, combining expert medical knowledge and a targeted focus on specific error types. This significantly enhances the overall accuracy and efficiency of the annotation process.

## 5 Conclusion

We introduce a novel approach, KeyBot, specifically designed to identify and correct significant errors in vertebrae keypoint estimation. Our proposed error simulation training procedure enables KeyBot to accurately and independently address specific error types, thereby significantly reducing the need for user intervention in the annotation process. Comprehensive evaluations on public real-world datasets confirm the superiority of KeyBot over existing methods.

## Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program (KAIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913).

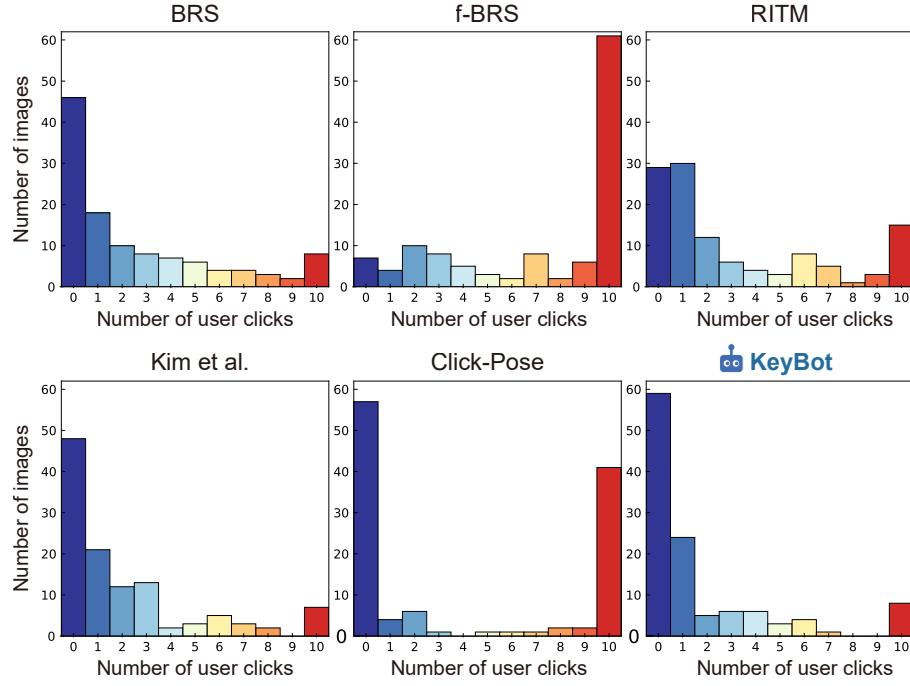
## References

1. Albiero, V., Chen, X., Yin, X., Pang, G., Hassner, T.: img2pose: Face alignment and detection via 6dof, face pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7617–7627 (2021)
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Colaco, S.J., Han, D.S.: Deep learning-based facial landmarks localization using compound scaling. *IEEE Access* **10**, 7653–7663 (2022)
4. Daniele, A., van Krieken, E., Serafini, L., van Harmelen, F.: Refining neural network predictions using background knowledge. *Machine Learning* pp. 1–39 (2023)
5. Guo, W., Du, Y., Shen, X., Lepetit, V., Alameda-Pineda, X., Moreno-Noguer, F.: Back to mlp: A simple baseline for human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4809–4819 (2023)
6. Guo, Y., Li, Y., Zhou, X., He, W.: A keypoint transformer to discover spine structure for cobb angle estimation. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
8. Jang, W.D., Kim, C.S.: Interactive image segmentation via backpropagating refinement scheme. *CVPR* pp. 5297–5306 (2019)
9. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 196–214. Springer (2020)
10. Kim, J., Kim, T., Kim, T., Choo, J., Kim, D.W., Ahn, B., Song, I.S., Kim, Y.J.: Morphology-aware interactive keypoint estimation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 675–685. Springer (2022)
11. Klinwichit, P., Yookwan, W., Limchareon, S., Chinnasarn, K., Jang, J.S., Onuean, A.: Buu-lspine: A thai open lumbar spine dataset for spondylolisthesis detection. *Applied Sciences* **13**(15), 8646 (2023)
12. Liu, Q., Zheng, M., Planche, B., Karanam, S., Chen, T., Niethammer, M., Wu, Z.: Pseudoclick: Interactive image segmentation with click imitation. In: European Conference on Computer Vision. pp. 728–745. Springer (2022)
13. Liu, T., Wang, Y., Yang, Y., Sun, M., Fan, W., Bunger, C., Wu, C.: A multi-scale keypoint estimation network with self-supervision for spinal curvature assessment of idiopathic scoliosis from the imperfect dataset. *Artificial Intelligence in Medicine* **125**, 102235 (2022)
14. Moon, G., Chang, J.Y., Lee, K.M.: Posefix: Model-agnostic general human pose refinement network. *CVPR* pp. 7773–7781 (2019)
15. Pisov, M., Kondratenko, V., Zakharov, A., Petraikin, A., Gombolevskiy, V., Morozov, S., Belyaev, M.: Keypoints localization for joint vertebra detection and fracture severity quantification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23. pp. 723–732. Springer (2020)
16. Qu, C., Zhang, T., Qiao, H., Tang, Y., Yuille, A.L., Zhou, Z., et al.: Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems* **36** (2024)

17. Rahmani, W., Suzuki, K., Lin, T.L.: Auto-ca: automated cobb angle measurement based on vertebrae detection for assessment of spinal curvature deformity. *IEEE Transactions on Biomedical Engineering* (2023)
18. Ruggero Ronchi, M., Perona, P.: Benchmarking and error diagnosis in multi-instance pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 369–378 (2017)
19. Shi, D., Wei, X., Li, L., Ren, Y., Tan, W.: End-to-end multi-person pose estimation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11069–11078 (2022)
20. Singh, S., Sandhu, N., Puri, T., Gulati, R., Kashyap, R.: A study of correlation of various growth indicators with chronological age. *International Journal of Clinical Pediatric Dentistry* **8**(3), 190 (2015)
21. Sofiuk, K., Petrov, I., Barinova, O., Konushin, A.: f-brs: Rethinking backpropagating refinement for interactive segmentation. *CVPR* pp. 8623–8632 (2020)
22. Sofiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3141–3145. IEEE (2022)
23. Türköz, Ç., Kaygısız, E., Ulusoy, Ç., Ateş, C.: A practical formula for determining growth. *Diagnostic and Interventional Radiology* **23**(3), 194 (2017)
24. Wu, H., Bailey, C., Rasoulinejad, P., Li, S.: Automatic landmark estimation for adolescent idiopathic scoliosis assessment using boostnet. In: MICCAI. pp. 127–135. Springer (2017)
25. Xu, Y., Xu, D., Hong, X., Ouyang, W., Ji, R., Xu, M., Zhao, G.: Structured modeling of joint deep feature and prediction refinement for salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3789–3798 (2019)
26. Yang, J., Zeng, A., Li, F., Liu, S., Zhang, R., Zhang, L.: Neural interactive key-point detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15122–15132 (2023)
27. Zhang, H., Chung, A.C.: A dual coordinate system vertebra landmark detection network with sparse-to-dense vertebral line interpolation. *Bioengineering* **11**(1), 101 (2024)

# Supplementary Materials for Bones Can't Be Triangles: Accurate and Efficient Vertebrae Keypoint Estimation through Collaborative Error Revision

**Summary.** This supplementary material enriches the main manuscript by providing comprehensive details of our methodology, additional visualizations, and extended experimental results. Section A presents additional experimental analyses, including a study of user click distribution, the integration of KeyBot with various interaction models, training KeyBot with real keypoint errors, annotation time comparison, and sensitivity analysis. Section B offers an in-depth description of our approach, outlining both its conceptual framework and algorithm. Section C elaborates on the implementation details, experimental setups, and baselines used for comparison. Section D investigates the application of KeyBot in a context where multiple refinement paths are explored, and the best path is chosen by the user, demonstrating its enhanced utility and effectiveness in



**Fig. 9:** Distribution of the number of user clicks (NoC) necessary to reach a target mean radial error (MRE) of 20 on the AASCE dataset.

the keypoint annotation process. Section E discusses the limitations and broad impact of our work. Section F provides additional qualitative results, further affirming the practicality and impact of our work.

## A More experimental results

Section A.1 analyzes the distribution of the number of user clicks. Section A.2 explores the integration of KeyBot with various interaction models. Section A.3 investigates training KeyBot with real keypoint errors. Section A.5 conducts a sensitivity analysis on the detector.

### A.1 Comparative analysis of user clicks required for target accuracy

In this analysis, we examine the distribution of the number of user clicks (NoC) required to achieve a target mean radial error (MRE) of 20, denoted as NoC<sub>10</sub>@20. We assess the performance of our proposed method, KeyBot, which operates with a maximum of three iterations, in comparison to various baseline models, as shown in Fig. 9. Notably, KeyBot reaches the target MRE with zero user clicks for a significant proportion of images, outperforming the baseline models. In more than half of the entire instances, the images achieve the required accuracy autonomously, demonstrating KeyBot’s efficiency in preemptively correcting major errors. KeyBot predominantly achieves the target MRE with fewer user interactions, reinforcing its effectiveness in reducing user effort while maintaining high accuracy in vertebrae keypoint estimation.

### A.2 Enhancing interactive keypoint estimation with keyBot

We rigorously assess KeyBot by integrating it with two interactive keypoint estimation frameworks, the model proposed by Kim et al. [10] and Click-Pose [26]. The results, shown in Table 5, include:

**Manual revision (manual).** Manual correction refers to user adjustments on the initial predictions of Kim et al. and Click-Pose without any subsequent model modification, assessing error reduction achieved solely through user revision.

**Model revision.** The model revision results in gray show improvements due to automated refinements based on user feedback. The difference from manual revisions quantifies the error reduction attributable to the interaction model.

**KeyBot.** The KeyBot results in blue demonstrate its performance when added to existing models. The same KeyBot instance integrates seamlessly with either framework, requiring no additional training.

**KeyBot without accumulating false predictions (w/o fp).** The results show the performance of KeyBot without the proposed false prediction accumulation strategy, specifically with the Kim et al. model, because this aspect is not applicable to Click-Pose (refer to Section C.4 for more details).

Our findings reveal that adopting KeyBot consistently improves the keypoint estimation performance of baseline models across all three datasets, demonstrating its robustness and effectiveness. KeyBot significantly reduces errors in

**Table 5:** Performance comparison of mean radial error in keypoint estimation across the AASCE, BUU-AP, and BUU-LA datasets. UC denotes the count of user clicks provided to the model.

Method	Interaction   backbone	AASCE				BUU-AP			BUU-LA			
		-	UC1	UC2	UC3	UC4	-	UC1	UC2	-	UC1	UC2
Kim et al. [10]	manual	51.58	49.73	48.07	46.49	44.98	42.31	38.65	35.48	23.43	20.81	18.63
Kim et al. [10]	Kim et al.	51.58	30.60	25.78	21.60	19.08	42.31	23.26	16.46	23.43	14.29	10.29
KeyBot-i3 w/o fp	Kim et al.	43.54	27.15	23.36	19.34	16.63	<b>31.85</b>	<b>20.65</b>	15.87	18.97	<b>13.34</b>	9.00
KeyBot-i1	Kim et al.	44.18	<b>25.93</b>	<b>20.66</b>	<b>18.03</b>	16.37	32.01	21.79	15.97	18.77	13.39	9.12
KeyBot-i2	Kim et al.	42.52	27.03	22.95	18.72	<b>16.23</b>	31.88	<b>20.65</b>	<b>15.81</b>	19.11	13.47	9.03
KeyBot-i3	Kim et al.	<b>41.70</b>	26.59	25.02	21.23	16.76	31.87	20.66	15.84	<b>18.74</b>	13.36	<b>8.97</b>
Click-Pose [26]	manual	54.65	52.80	51.27	49.85	48.48	32.72	29.25	26.13	33.70	30.02	26.88
Click-Pose [26]	Click-Pose	54.65	46.50	44.08	41.73	40.04	32.72	29.30	26.20	33.70	21.62	17.38
KeyBot-i1	Click-Pose	52.62	46.10	43.65	41.17	39.39	31.66	28.03	24.50	33.98	20.53	<b>16.87</b>
KeyBot-i2	Click-Pose	51.38	45.98	43.53	40.98	39.09	<b>31.45</b>	<b>27.91</b>	<b>24.27</b>	33.29	20.31	16.97
KeyBot-i3	Click-Pose	<b>51.24</b>	<b>45.92</b>	<b>43.48</b>	<b>40.89</b>	<b>38.87</b>	31.56	28.01	24.35	<b>32.39</b>	<b>20.19</b>	17.14

**Table 6:** Comparison of real and synthetic errors on the AASCE dataset.

Training	Mean radial error				NoC <sub>10</sub>					
	-	UC1	UC2	UC3	UC4	@20	@30	@40	@50	
syn	X	51.58	30.60	25.78	21.60	19.08	2.10	1.54	1.23	1.03
real	✓	<b>41.70</b>	<b>26.59</b>	<b>25.02</b>	21.23	<b>16.76</b>	<b>1.74</b>	<b>1.32</b>	<b>0.94</b>	<b>0.68</b>
	✓	46.42	28.64	26.09	<b>19.89</b>	17.80	1.88	1.34	0.99	0.72
	X	51.55	30.64	26.44	21.14	18.37	2.11	1.52	1.28	1.01

scenarios with no user interaction. For instance, on the AASCE dataset, KeyBot notably lowers MREs, a trend also observed in the BUU-AP and BUU-LA datasets. The model demonstrates further error reduction post user interactions. Its variant (w/o fp) exhibits slightly lower, yet still notable, performance. Overall, the results demonstrate that KeyBot is model-agnostic and can be adapted to different frameworks without the need for model-specific training.

### A.3 Training KeyBot with real keypoint errors

We experiment with including real keypoint mistakes in the KeyBot training dataset, using a probability distribution of 40% real mistakes, 40% synthetic errors, and 20% accurate keypoints. However, including real errors resulted in decreased performance, as shown in Table 6. Real errors have high variability and lack consistent patterns, making it challenging to identify a clear pattern to correct. In contrast, synthetic errors are clearly defined and consistent, facilitating better convergence during training and leading to improved performance.

**Table 7:** Comparison of annotation cost per image on the BUU-LA dataset.

Method	Time (s)	User click
manual revision	$23.10 \pm 2.93$	$7.02 \pm 1.74$
model revision w/o KeyBot	$21.61 \pm 12.66$	$3.04 \pm 1.94$
model revision w/ KeyBot	$11.71 \pm 2.23$	$0.44 \pm 0.30$

**Table 8:** Sensitivity analysis of the detector. MRE is measured on the BUU-AP and BUU-LA datasets.  $k$  and  $s$  denotes represents the number of keypoints examined simultaneously and the stride during inference, respectively. UC denotes the number of user clicks. KeyBot-i3 is used for the analysis.

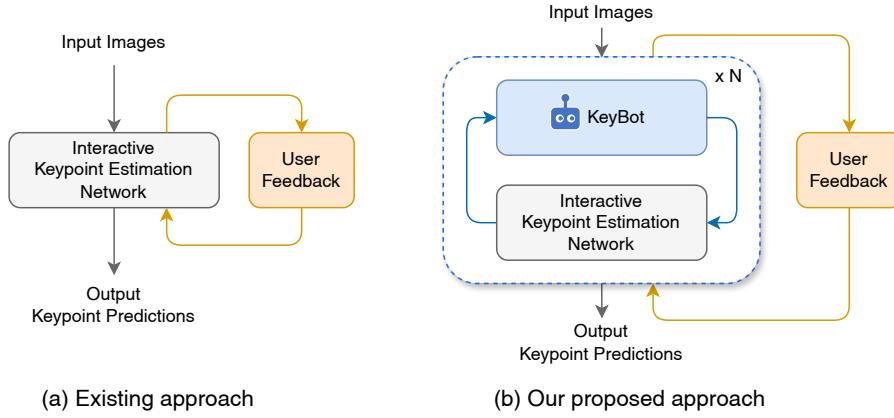
Detector	BUU-AP						BUU-LA		
	$k$	$s$	-	UC1	UC2	-	UC1	UC2	
	68	-	39.32	22.39	19.64	20.44	13.94	9.53	
	8	1	31.87	18.30	16.02	<b>18.61</b>	14.87	<b>8.94</b>	
	8	2	34.85	18.70	15.97	18.64	13.44	8.96	
	8	3	<b>31.75</b>	<b>18.25</b>	15.94	18.74	13.44	8.99	
	8	4	31.87	20.66	<b>15.84</b>	18.74	<b>13.36</b>	8.97	
	4	1	36.17	20.25	16.24	18.79	13.42	8.99	
	4	2	32.22	20.86	16.34	<b>18.42</b>	<b>13.39</b>	9.07	
	4	3	<b>31.74</b>	<b>18.14</b>	<b>16.03</b>	18.56	13.46	10.07	
	4	4	32.27	22.94	16.14	18.61	15.10	<b>8.95</b>	

#### A.4 Comparison of annotation time

We conduct a user study with 15 participants, each tasked with annotating 22 keypoints on ten challenging radiographs from the BUU-LA dataset [11]. Participants are divided into three groups: one using KeyBot, one without it, and one using only initial model predictions without subsequent model revision. As summarized in Table 7, the results show that KeyBot significantly reduces annotation time and user clicks, demonstrating its efficiency. Although the computation time for KeyBot is higher, its average inference time remains under 0.22 seconds, which is negligible.

#### A.5 Sensitivity analysis of the detector in keypoint error detection

We conduct a sensitivity analysis on the detector, as detailed in Table 8. We investigate the impact of varying the number of simultaneously examined keypoints ( $k$ ) and the stride ( $s$ ) during inference on keypoint estimation accuracy. We observe a substantial decrease in performance when the detector assesses all keypoints at once, resulting in the most significant keypoint estimation errors. This performance decline suggests that analyzing the entire bone structure



**Fig. 10:** Comparison between the (a) existing interactive keypoint estimation framework and (b) the proposed approach adopting KeyBot.

at once increases complexity, diminishing the detector's effectiveness. Focusing on specific bone segments during detection proves to be more efficient. KeyBot maintains robust performance with different input keypoint numbers, especially when  $k = 4$  and  $k = 8$ .

Additionally, we analyze the effect of varying the stride  $s$  with a fixed keypoint window  $k$ . KeyBot performs consistently across different stride settings. Particularly with an input keypoint number of four, smaller stride values than  $k$  slightly enhance performance. A smaller stride provides a more detailed and contextually varied examination of each keypoint, improving the detector's accuracy and reliability.

## B Additional details of KeyBot

This section complements an explanation of KeyBot. First, Section B.1 describes the detailed conceptual framework of our method, and Section B.2 offers an algorithm that elucidates the overall process of KeyBot.

### B.1 Overview of the proposed approach

Existing interactive keypoint estimation approaches [10, 26] operate by predicting keypoints from images and refining these predictions based on user feedback, as depicted in Fig. 10(a). However, they lack the capability for self-correction without user intervention.

To address this limitation, we introduce KeyBot, which independently evaluates the interaction model's predictions, pinpointing and correcting errors autonomously, as illustrated in Fig. 10(b). This approach facilitates autonomous correction of predictions, thereby enabling users to concentrate on subsequent refinements after KeyBot's preliminary error rectification.

**Why a user-interactive approach?** In medical imaging, the accuracy of keypoint estimation models is paramount. Despite recent advancements, human adjustments are often needed to address inherent model biases and errors. However, manual revision, while necessary to ensure reliability and accuracy, is often a tedious and time-consuming task. Thus, an interactive approach incorporating human feedback aims to streamline this process by refining inaccuracies with minimal user intervention.

**Why KeyBot?** Existing interactive models require users to initiate error assessment, which can be laborious and error-prone, especially with numerous predictions or clustered inaccuracies. KeyBot autonomously conducts an initial assessment, addressing basic errors such as fundamental misidentifications of bone structures. By addressing these primary errors, it reduces the time users spend on basic error corrections, allowing for a more focused and efficient use of human expertise.

**Why not end-to-end?** While an interactive keypoint estimation model learns to correct its errors based on user feedback, it lacks explicit training for specific error types. In contrast, KeyBot is designed to identify and correct three specific error types, utilizing error simulation in its training phase. Its independent structure prevents it from inheriting potential biases or limitations of the interaction model, ensuring an objective and reliable keypoint estimation process.

**Ongoing collaborative loop** Integrating KeyBot within the interactive keypoint estimation framework in a feedback-providing manner maintains an iterative and collaborative loop. This loop ensures that the refinement process is not a one-off correction but a continuous improvement cycle, accommodating further refinements and adjustments from both KeyBot and human users, leading to more accurate outcomes in medical image analysis.

## B.2 Algorithm of KeyBot

The complete procedure of our proposed approach is encapsulated in Algorithm 1. Given an input image, the interaction model makes an initial keypoint prediction, followed by two sequential phases: the KeyBot phase (with  $N$  iterations) and the user phase (one iteration). During the KeyBot phase, KeyBot performs a preliminary step to correct errors and generates corrective feedback. This feedback is fed into the interaction model, akin to user feedback. The interaction model then generates a new, corrected prediction. This phase repeats for  $N$  iterations or until KeyBot detects no errors. Subsequently, in the user phase, the user corrects a single error, and the interaction model updates its results accordingly. This entire process repeats until it reaches the maximum number of user interactions,  $T$ .

**Algorithm 1:** Inference with KeyBot

---

**Data:** input image  $\mathbf{x}$ , maximum user click  $T$ , maximum KeyBot iteration number  $N$ , interaction model  $\mathcal{F}_\theta$ , KeyBot detector  $\mathcal{G}_{\phi_r}$ , KeyBot corrector  $\mathcal{G}_{\phi_s}$

**Result:**  $\mathbf{y}$ , the final keypoint prediction

$$\mathbf{c}_{0,0} = 0, \mathbf{e}_{0,0} = 0$$

$$\mathbf{y}_{0,0} \leftarrow \mathcal{F}_\theta(\mathbf{x}, \mathbf{c}_{0,0}, \mathbf{e}_{0,0}) \quad \triangleright \text{initial interaction model forward}$$

```

while  $t < T$  do
    while  $n < N$  do
         $\nu_{t,n+1} \leftarrow \mathcal{G}_{\phi_r}(\mathbf{x}, \mathbf{y}_{t,n}) \quad \triangleright \text{KeyBot detector}$ 
        if  $|\nu_{t,n+1} \setminus \rho_t| = 0$  then
            break
        else
             $\mathbf{z}_{t,n+1} \leftarrow \mathcal{G}_{\phi_s}(\mathbf{x}, \mathbf{y}_{t,n}) \quad \triangleright \text{KeyBot corrector}$ 
             $\mathbf{c}_{t,n+1}^i \leftarrow \mathbf{z}_{t,n+1}^i, i \in \nu_{t,n+1} \setminus \rho_t$ 
             $\mathbf{e}_{t,n+1}^i \leftarrow \mathbf{y}_{t,n}^i, i \in \nu_{t,n+1} \setminus \nu_{t,n} \setminus \rho_t$ 
             $\mathbf{y}_{t,n+1} \leftarrow \mathcal{F}_\theta(\mathbf{x}, \mathbf{c}_{t,n+1}, \mathbf{e}_{t,n+1}) \quad \triangleright \text{interaction model forward}$ 
        end
    end
     $\mathbf{c}_{t+1,0} \leftarrow 0, \mathbf{e}_{t+1,0} \leftarrow \mathbf{e}_{t,n+1}$ 
 $\rho_{t+1}, \mathbf{u}_{t+1} \leftarrow \Psi(\mathbf{x}, \mathbf{y}_{t,n+1}, \mathbf{u}_t) \quad \triangleright \text{User revision}$ 
 $\mathbf{c}_{t+1,0}^i \leftarrow \mathbf{u}_{t+1}^i, i \in \rho_{t+1}$ 
 $\mathbf{e}_{t+1,0}^i \leftarrow \mathbf{y}_{t,n+1}^i, i \in \rho_{t+1}$ 
 $\mathbf{y}_{t+1,0} \leftarrow \mathcal{F}_\theta(\mathbf{x}, \mathbf{c}_{t+1,0}, \mathbf{e}_{t+1,0}) \quad \triangleright \text{interaction model forward}$ 
end

```

---

## C Experimental details

This section provides comprehensive details about the implementation details (Section C.1), dataset descriptions (Section C.2), metric definitions (Section C.3), and reproducibility details for baseline models (Section C.4).

### C.1 Implementation details of KeyBot

We describe a detailed experimental setup of our approach: the detector, the corrector, and the interaction model. Also, we elaborate on the error simulation methods employed in training the detector and the corrector, including vertex misidentification (misvertex), bone misidentification (misbone), and left-right inversion (lr-inversion).

**Detector** The detector analyzes eight ( $k = 8$ ) keypoints simultaneously, classifying each as accurate or inaccurate. Input X-ray images, cropped and resized

around these keypoints to  $128 \times 128$  dimensions, are concatenated with Gaussian keypoint heatmaps. The detector evaluates the abnormality likelihood for each keypoint using a sigmoid function. Training labels are generated by marking synthetically displaced keypoints as one (indicative of errors) and the rest as zero. During inference, the detector iteratively processes keypoints with a stride of four ( $s = 4$ ). On the BUU-LA datasets, this process is applied to only 20 keypoints, excluding the final two for error detection. Any keypoint with an anomaly probability above 0.5 is flagged as erroneous.

The detector employs a modified ResNet-18 [7] architecture, adapted for combined image and heatmap inputs. The training process incorporates simulated vertex misidentification errors, displacing up to three keypoints per image for AASCE and four keypoint for BUU-AP and BUU-LA datasets, respectively. Selected keypoints shift up to four indices away from their original position, with wrapping for out-of-range indices. The training utilizes Binary Cross-Entropy (BCE) loss over 300 epochs with early stopping (zero patience) and an AdamW optimizer with a learning rate of 0.001.

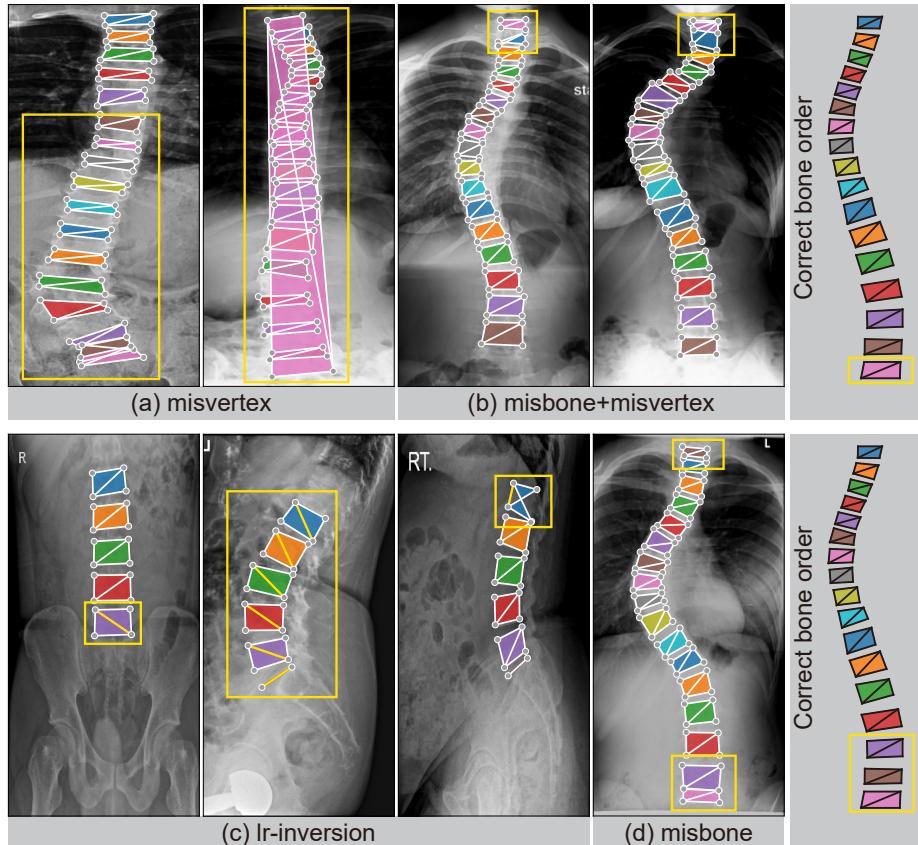
**Corrector** The corrector processes the entire image, resized to  $256 \times 128$ , alongside Gaussian keypoint heatmaps of matching resolution. It generates reconstructed keypoint locations as heatmaps, using a sigmoid function in the final layer. The architecture is based on DeepLab-v3 with a ResNet50 encoder. During training, KeyBot is trained on accurate keypoints with a 20% probability and simulated errors with an 80% probability. Training uses three error types or accurate keypoints, with varied probabilities for each dataset (Please refer to the source code for more details).

- (1) **Vertex misidentification errors:** Up to nine keypoints are displaced, with a multinomial probability distribution for the number of keypoints to shift. Keypoints are selected randomly with equal probability, and they are shifted to maximally four indices away from their original index.
- (2) **Bone misidentification errors:** To simulate misbone errors, movement type (up, down, or accurate) is selected with equal probability. For the shifts, there is an equal probability for each of the following scenarios: moving all keypoints from the first to the last, moving keypoints from a random starting point to the last keypoint, moving keypoints from the first to a random end point, and moving keypoints between a random start and end point. For keypoints located at either the first or last vertebrae, relocation may occur outside the targeted bone structure, with the magnitude determined by positional differences between either the first and second vertebrae or the last and its immediate predecessor.
- (3) **Left-right inversion errors:** Every left-right pair is independently swapped with a 90% probability.

Similar to the detector, the corrector is trained over 300 epochs with early stopping, using an AdamW optimizer with the learning rate of 0.001.

**Table 9:** Summary of X-ray image datasets for keypoint estimation used in our work.

Dataset	Target keypoints	Number of images				Human annotation error			
		Total	Train	Val	Test	Total	misvertex	misbone	lr-inversion
AASCE [24]	68	563	325	122	116	45	18	27	-
BUU-AP [11]	20	399	240	80	79	1	-	-	1
BUU-LA [11]	22	397	237	80	80	3	-	-	3

**Fig. 11:** Errors in human annotation present in the AASCE dataset in (a),(b) and (d), and in the BUU-AP and BUU-LA datasets in (c). A diagonal line connecting the upper-right vertex to the lower-left vertex is indicated in each vertebra.

**Interaction model** The interaction model estimates keypoint heatmaps from input X-ray images with a size of  $512 \times 256$ . We extend the model proposed by Kim et al. [10] by integrating revision feedback from either KeyBot or user input and adopting accumulated false predictions. Training includes iterative

refinement based on simulated user feedback, providing groundtruth keypoint locations as user revision. During inference, the most significant keypoint error is modified to simulate user interaction.

### C.2 Datasets

In our study, we utilize public X-ray image datasets, namely AASCE, BUU-AP, and BUU-LA, as detailed in Table 9. An extensive analysis of these datasets reveals three primary types of human annotation errors: vertex misidentification (misvertex), bone misidentification (misbone), and left-right inversion (lr-inversion). The AASCE dataset predominantly exhibits misvertex and misbone errors, whereas the BUU-AP and BUU-LA datasets mainly exhibit lr-inversion errors. To ensure the integrity and reliability of our evaluation, images with such critical annotation errors were excluded, as illustrated in Fig. 11.

**AASCE [24]** The AASCE dataset consists of 608 anterior-posterior X-ray images, annotated with 68 keypoints across 17 vertebrae. We follow the original train-test split, further partitioning the training set for validation purposes. Error analysis identifies that it contains nine misvertex and 18 misbone errors in the training set, one misvertex and five misbone errors in the validation set, and eight misvertex and four misbone errors in the test set, as exemplified in Fig. 11(a,b,d). Two cases showing a mixture of misvertex and misbone errors are counted as misbone errors.

**BUU-AP [11]** Comprising 400 anterior-posterior view X-ray images, BUU-AP dataset is annotated with 20 keypoints per image. The image sizes range from  $1434 \times 1072$  to  $3072 \times 3040$ . We randomly split the images into training, validation, and test sets, identifying a lr-inversion error in the test set, as shown in Fig. 11(c).

**BUU-LA [11]** The BUU-LA dataset includes 400 left lateral view X-ray images, each with 22 annotated spinal keypoints with image sizes ranging from  $1956 \times 968$  to  $3072 \times 3040$ . It undergoes similar partitioning as BUU-AP, with three lr-inversion errors identified in the training set, as exemplified in Fig. 11(c).

### C.3 Definition of mean radial error

In our study, we assess the keypoint estimation accuracy using the mean radial error (MRE). MRE measures the average Euclidean distance between the predicted and groundtruth coordinates of keypoints. Specifically, for each of the  $K$  target keypoints in a sample, let  $\mathbf{p}_i^*$  denote the groundtruth coordinates and  $\mathbf{p}_i$  the predicted coordinates of the  $i$ -th keypoint. The MRE is calculated as:

$$\text{MRE} = \frac{1}{K} \sum_{i=1}^K \|\mathbf{p}_i^* - \mathbf{p}_i\|_2. \quad (1)$$

MRE measures the overall precision of keypoint estimation results.

#### C.4 Reproducibility for baselines

**Interactive segmentation models.** In this work, we compare our method with several interactive segmentation models, including BRS [8]<sup>1</sup>, f-BRS [21]<sup>2</sup>, and RITM [22]<sup>3</sup>, for our keypoint estimation task. Using their official source codes, we modify these models to produce outputs for the specific number of keypoints required in our study, aligning with the hyperparameter settings from Kim et al. [10] for consistency.

**Kim et al. [10]** We adhere to the experimental settings outlined in the official source code<sup>4</sup>. For experiments on the AASCE dataset, we utilize their pretrained network. In the case of the BUU-AP and BUU-LA datasets, we train the model using identical hyperparameters as those used for the AASCE dataset, with the only modification being the selection of keypoint subsets. This adjustment is necessary to accommodate the morphology-aware loss proposed in the study. When integrating KeyBot, we handle KeyBot’s revision feedback as user interaction input, similar to processing user clicks.

**Click-Pose [26]** We utilize the official source code of Click-Pose<sup>5</sup>, maintaining the hyperparameters consistent across all datasets, while adjusting the number of keypoints as required. During the training phase, we exclude the Object Keypoint Similarity (OKS) loss which is specifically tailored for human pose estimation. As a result, Click-Pose training involves using a combination of L1 loss for bounding boxes, intersection over union (IOU) loss, classification loss, and L1 loss for keypoints. Click-Pose employs a keypoint regression-based approach, directly estimating the coordinates of keypoints. For integration with KeyBot, we feed KeyBot’s revision feedback into the human-to-keypoint decoder module of Click-Pose, akin to handling user clicks. Given that Click-Pose does not employ a heatmap-based keypoint estimation approach, we do not incorporate false predictions into the input framework.

## D Exploring multiple refinement paths of KeyBot

This section introduces a novel collaborative annotation approach involving KeyBot, the user, and the interaction model. We investigate the application of KeyBot in a context where multiple refinement paths are explored, and the best path is chosen by the user, demonstrating its enhanced utility and effectiveness in the keypoint annotation process. This method involves multiple iterative interactions between KeyBot and the backbone model, yielding a variety of refined results. These results, alongside the initial prediction, are presented to the

---

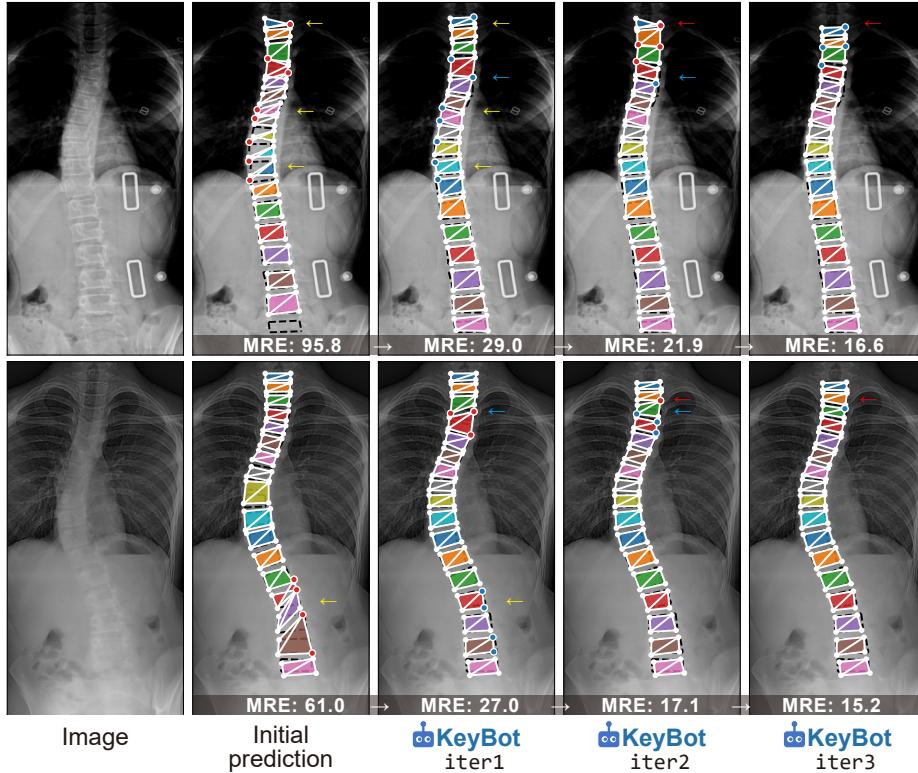
<sup>1</sup> [https://github.com/wdjang/BRS-Interactive\\_segmentation](https://github.com/wdjang/BRS-Interactive_segmentation)

<sup>2</sup> [https://github.com/saic-vul/fbrs\\_interactive\\_segmentation/tree/master](https://github.com/saic-vul/fbrs_interactive_segmentation/tree/master)

<sup>3</sup> [https://github.com/SamsungLabs/ritm\\_interactive\\_segmentation](https://github.com/SamsungLabs/ritm_interactive_segmentation)

<sup>4</sup> [https://github.com/seharanul17/interactive\\_keypoint\\_estimation](https://github.com/seharanul17/interactive_keypoint_estimation)

<sup>5</sup> <https://github.com/IDEA-Research/Click-Pose>



**Fig. 12:** Iterative refinement results of KeyBot on the AASCE dataset. In the first example, KeyBot corrects the upper and middle part in the image indicated by the yellow arrow, lowering the bones overall in the first iteration. It revises the blue arrow area in the second iteration and finally lowers the incorrectly positioned bones at the top in the third iteration. In the second example, the first iteration adjusts the lower part as indicated by the yellow arrow. In the second iteration, the front end of the bones marked in blue is lowered overall. Finally, in the third iteration, the model properly adjusts the position of the bones at the top, marked in red. A diagonal line connecting the upper-right vertex to the lower-left vertex is indicated in each vertebra.

user for selection and potential further refinement. Among multiple refinement iterations, users can select the most precise result as their foundation for any additional modifications. Users also have the option to ignore the model updates and correct them manually if needed.

This approach resembles the concept of beam search, where each KeyBot refinement iteration represents a node in the search space. We allow users to explore multiple paths, i.e., KeyBot iterations, and to choose the best path. Providing multiple iteration results is feasible because the average inference time for the interaction model is 0.181 seconds, and for KeyBot is 0.216 seconds, which is negligible. The procedure consists of three steps. Initially, the backbone model

**Table 10:** Performance comparison of mean radial error in keypoint estimation across the AASCE, BUU-AP, and BUU-LA datasets. For KeyBot-i3 w/ beam search, the best prediction is chosen among three iterations. No user input is provided.

Method	AASCE	BUU-AP	BUU-LA
BRS [8]	45.65	51.22	40.20
f-BRS [21]	64.06	44.05	36.03
RITM [22]	56.03	36.43	23.27
Kim et al. [10]	51.58	42.31	23.43
Click-Pose [26]	54.65	32.72	33.70
KeyBot-i1	44.18	32.01	18.77
KeyBot-i2	42.52	31.88	19.11
KeyBot-i3	41.70	31.87	18.74
KeyBot-i3 w/ beam search	<b>39.29</b>	<b>31.43</b>	<b>18.70</b>

produces an initial keypoint prediction from the input X-ray image, forming the baseline prediction. Next, KeyBot evaluates this prediction, identifies inaccuracies, and makes corrections, leading to the first refined set of keypoints. This iteration repeats, resulting in multiple refined sets. Lastly, the user reviews all keypoint sets, including the initial and refined ones, and selects the most accurate set as a basis for any further adjustments.

The primary benefit of this approach is that it provides users with multiple refined results in addition to the initial prediction, as shown in Fig. 12. This allows for a well-informed decision-making process, where users select the most precise result as their foundation for any additional modifications.

Additionally, we analyze the error reduction achieved through this multi-iteration presentation approach, as shown in Table 10. This multi-iteration presentation approach with KeyBot, combining KeyBot’s iterations with user selection and refinement, optimizes prediction accuracy while significantly reducing user effort.

## E Discussion

The anatomical complexity and similarity among vertebrae make vertebrae keypoint estimation prone to errors that require substantial human efforts to correct. By advancing methodology to enhance accuracy and efficiency in this area, our work represents a substantial technical contribution. However, a limitation is that its correction precision occasionally falls short of human-level precision when most keypoints contain substantial errors, as shown in Fig. 8 of our main manuscript. In these cases, KeyBot disregards the erroneous predictions and generates completely new ones. However, because KeyBot generates revisions based on the initial predictions (refer to Fig. 4 of our main manuscript), it is

influenced by these errors, limiting its ability to provide precise feedback for incorrect vertebrae shapes. This necessitates user feedback to guide further precise adjustments. Future work aims to develop a more robust feedback mechanism to address these cases effectively, enhance accuracy, and minimize user input.

Although our approach is developed specifically for vertebrae, it introduces a general framework for addressing domain-specific errors that can be extended to other fields. By characterizing domain-specific error types and generating synthetic data to represent these inaccuracies, our approach can be used to develop an auxiliary model to detect and correct them efficiently. This adaptability highlights the broader impact and versatility of our work.

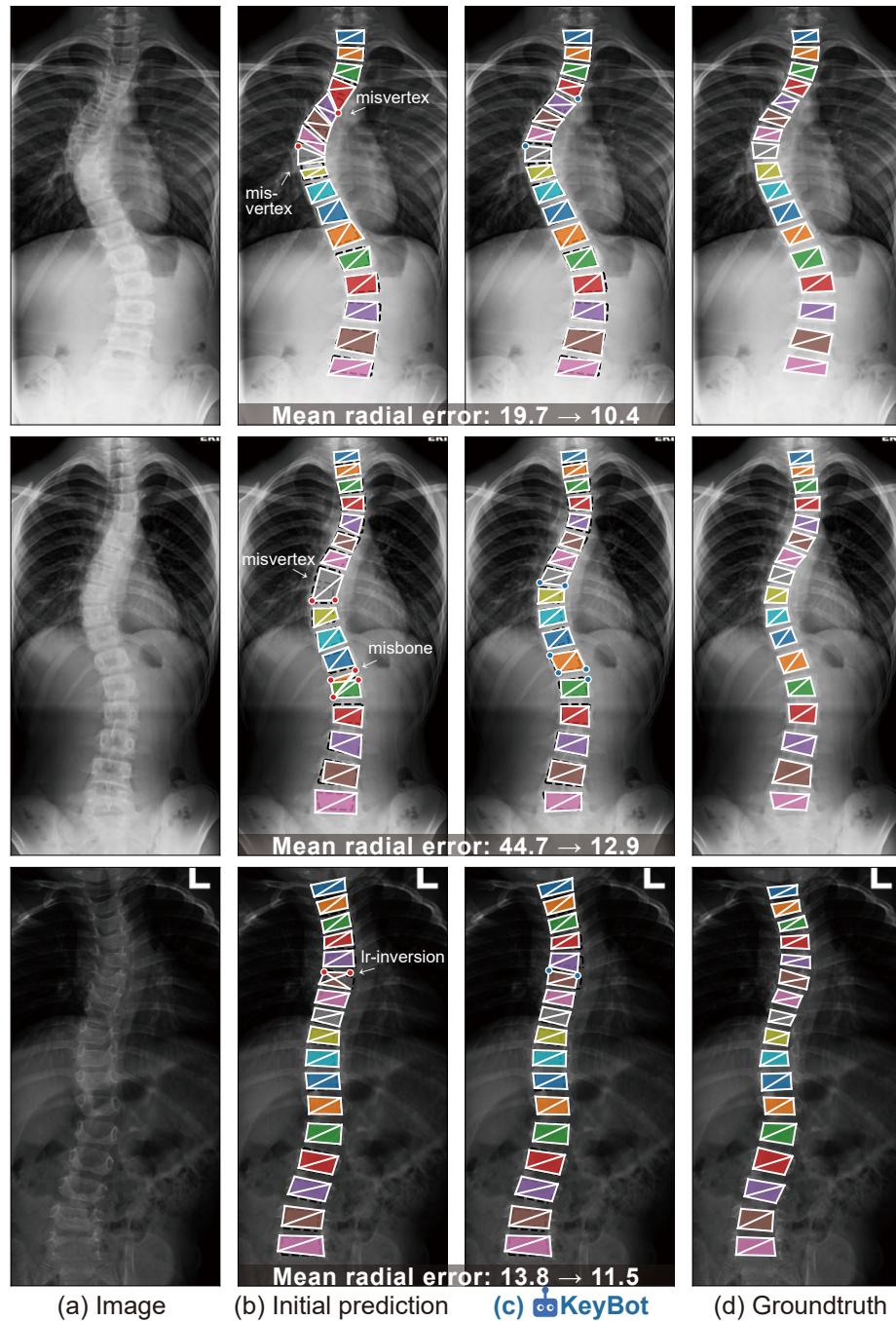
## F Additional qualitative results

This section presents additional qualitative results on the AASCE dataset, as depicted in Figs. 13 and 14, and on the BUU-AP and BUU-LA datasets, as shown in Fig. 15, highlighting the effectiveness of KeyBot in autonomously revising multiple keypoints simultaneously, minimizing the need for user intervention. A diagonal line connecting the upper-right vertex to the lower-left vertex is indicated in each vertebra.

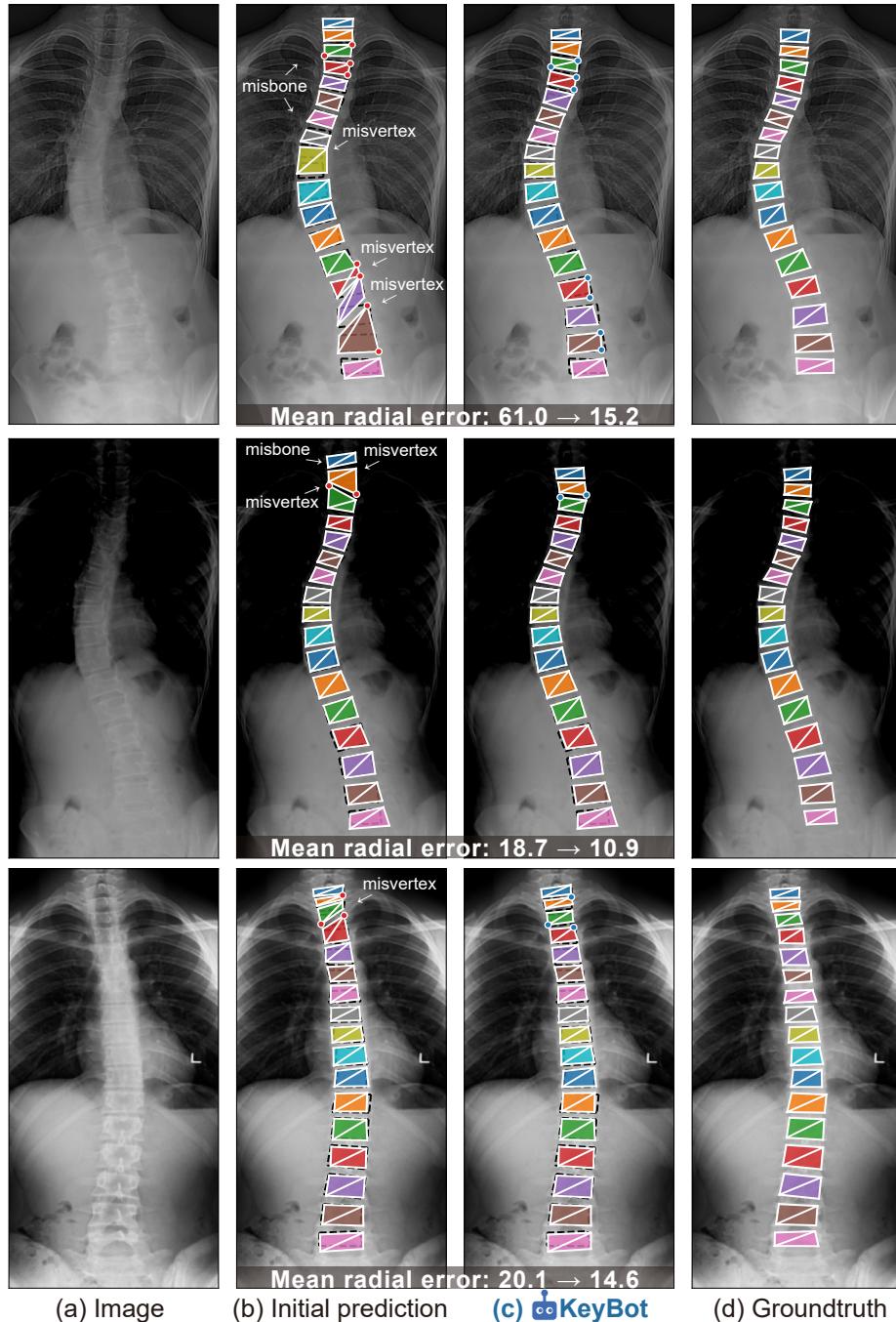
Initial model predictions often exhibit significant inaccuracies in representing bone shape, posing challenges for users in identifying and correcting errors, mainly when keypoints are densely clustered or overlap. KeyBot significantly streamlines this error correction process by effectively identifying erroneous keypoints, which are marked with red dots in the initial predictions. It distinguishes well-represented bone shape keypoints from severe morphological distortions, including misvertex errors (incorrect positioning of a portion of vertebra keypoints) and lr-inversion errors (incorrect left-right orientation). KeyBot is also highly effective in recognizing misbone errors, where an entire bone is misidentified, as illustrated in Figs. 13 and 14.

These capabilities allow KeyBot to provide targeted interventions for specific errors, greatly enhancing the accuracy and reliability of the overall keypoint estimation process. By automating the identification and correction of such errors, KeyBot not only improves the precision of the model but also reduces the burden on users, minimizing the cognitive load on users, allowing them to concentrate on verifying and fine-tuning the results, requiring considerably less effort.

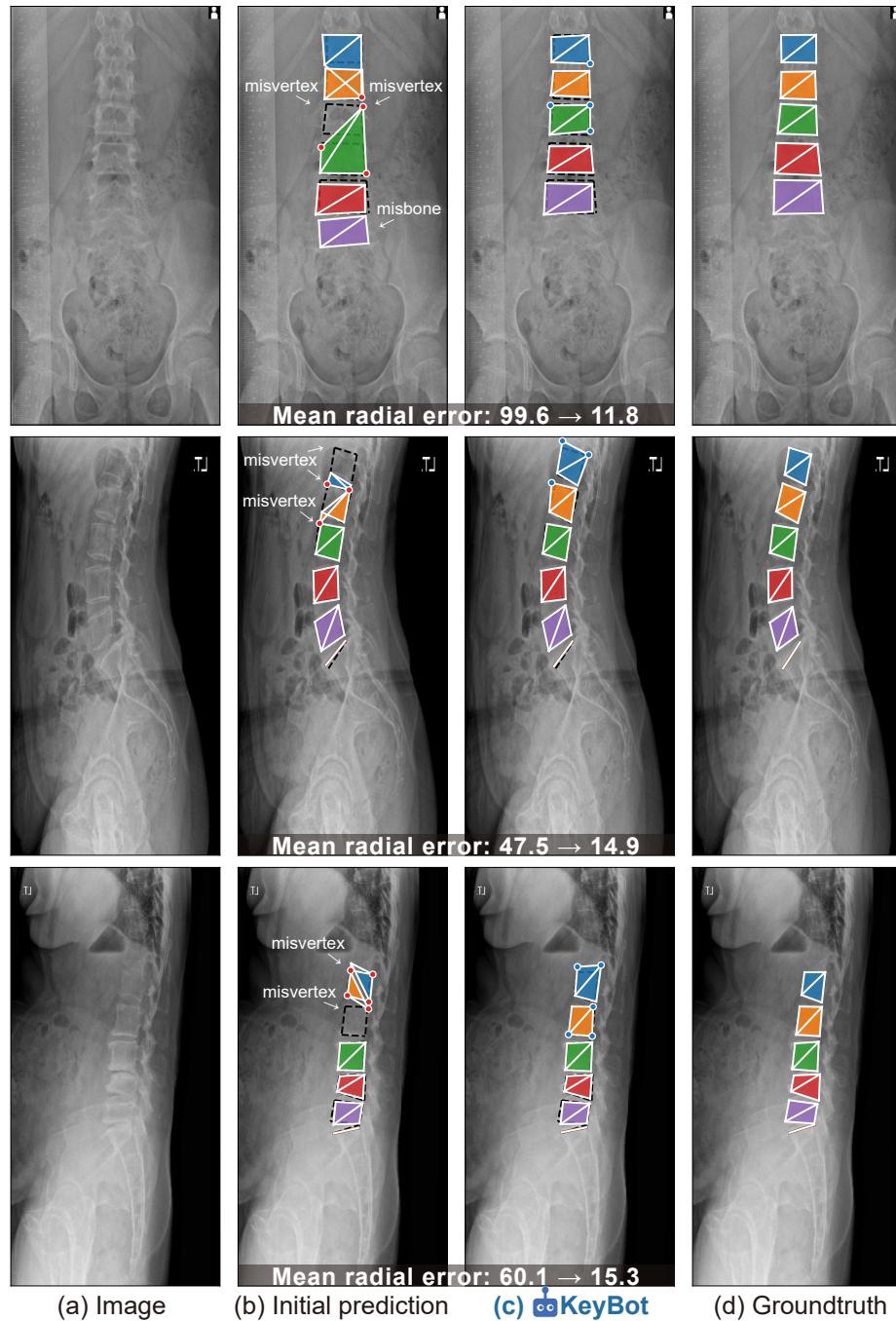
For a more comprehensive understanding of KeyBot’s capabilities, please refer to our demo video, which visually represents KeyBot’s efficiency and accuracy across various examples.



**Fig. 13:** Additional qualitative results of KeyBot on the AASCE dataset, with a maximum of three iterations.



**Fig. 14:** Additional qualitative results of KeyBot on the AASCE dataset, with a maximum of three iterations.



**Fig. 15:** Additional qualitative results of KeyBot on the BUU-AP (top) and BUU-LA (middle and bottom) datasets, with a maximum of three iterations.