X2Bone: Reconstructing 3D Bone Structures from 2D Biplanar X-Rays via Cross-RWKV

Zhaohong Pan[1,2], Jiahua Zhu[1,2], Haowei Zhou[1,2], Guang Yang[1,2], Jingjing Dai[1,2], Yaoqin Xie[1], and Xiaokun Liang[1⋆]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2] University of Chinese Academy of Sciences, Beijing, China
xk.liang@siat.ac.cn

Abstract. Reconstructing patient-specific 3D pelvic bone structures from biplanar X-ray images is attractive for routine clinical workflows, as it enables 3D understanding from widely available low-dose radiographs. Existing methods often lack effective cross-view interaction to fully exploit the complementarity between anterior–posterior (AP) and lateral (LAT) views, and provide limited global structural dependency modeling, which leads to structural ambiguity and missing details under limited projections. To address these issues, we propose X2Bone, an efficient end-to-end framework for 3D bone reconstruction from standard AP and LAT radiographs. X2Bone introduces a Cross-RWKV module that enables long-range, bidirectional cross-view interaction with linear complexity, achieving global semantic alignment and feature fusion across views. To bridge the 2D-to-3D representation gap, we further design a 3D Expand module, which lifts the fused 2D features into volumetric space via view-dependent depth probability modulation. Extensive experiments on the CTPelvic1K dataset demonstrate that X2Bone achieves the best segmentation performance while maintaining high efficiency. The results validate the effectiveness of long-range cross-view modeling and view-aware 2D-to-3D lifting for biplanar 3D reconstruction. The code is available at https://github.com/pzhhhhh2263/X2Bone.

Keywords: Reconstruction · Biplanar X-Rays · Cross-RWKV · 2D-3D

1 Introduction

Accurate 3D bone geometry is crucial for orthopedic diagnosis, preoperative planning, and computer-assisted surgery, particularly in total hip arthroplasty [11,13,17]. Although computed tomography (CT) provides high-fidelity patient-specific 3D anatomy, its relatively high radiation dose, cost, and reliance on specialized equipment limit routine use, especially for large-scale screening and resource-constrained settings [6].

⋆ Corresponding author.

X-ray imaging, in contrast, is widely available, inexpensive, and involves substantially lower radiation exposure [16]. This motivates reconstructing 3D bone anatomy from a small number of 2D radiographs, which is clinically attractive yet inherently ill-posed [20]. In practice, anterior–posterior (AP) and lateral (LAT) radiographs are the standard biplanar acquisition. However, severe depth ambiguity and anatomical overlap in 2D projections make accurate recovery of 3D geometry highly challenging. Earlier statistical shape model methods can incorporate anatomical priors, but typically require complex pre-modeling and careful initialization, and are sensitive to modeling assumptions [1,9,14].

Recent deep learning approaches enable end-to-end biplanar 3D reconstruction by learning direct mappings from AP/LAT images to volumetric, point-based, or mesh-based 3D representations, commonly using encoder–decoder frameworks with feature fusion [20]. Despite progress, performance under limited-view settings is still constrained by two key factors: (i) insufficient modeling of long-range structural dependencies, and (ii) ineffective cross-view interaction that fails to fully exploit AP/LAT complementarity.

Convolutional neural network models are limited by their effective receptive fields, while transformer-based designs improve global context modeling but suffer from quadratic complexity with respect to spatial resolution, which hampers high-resolution reconstruction. Meanwhile, existing pseudo-3D reshaping or input-level fusion strategies often provide weak and unstable cross-view coupling, leaving substantial structural ambiguity unresolved.

Against this background, achieving efficient modeling of long-range dependencies while fully exploiting cross-view complementarity remains a central challenge in biplanar 3D bone reconstruction. Recently, the Receptance Weighted Key–Value (RWKV) model [4,18] reformulates attention as a recurrent computation with linear complexity, enabling effective long-range dependency modeling without explicitly constructing global attention matrices. This property offers a promising direction for efficient global modeling [5,22,23]. However, its extension to cross-view interaction modeling in biplanar medical imaging remains largely unexplored.

To address these challenges, we propose a novel framework for 3D bone reconstruction from biplanar X-ray images. First, we design a Cross-RWKV interaction mechanism that explicitly enables global semantic alignment and fusion between AP and LAT feature streams while preserving the linear-complexity advantage of RWKV. Second, to mitigate the dimensional gap and information loss during the 2D-to-3D transformation, we introduce a depth-attention-based 2D–3D Lifting Module, which performs information-preserving volumetric lifting via orthogonal back-projection.

In summary, the main contributions of this work are as follows:

1. We propose an efficient biplanar X-ray-based 3D reconstruction framework with explicit cross-view global modeling, achieving significantly improved reconstruction accuracy over existing methods.

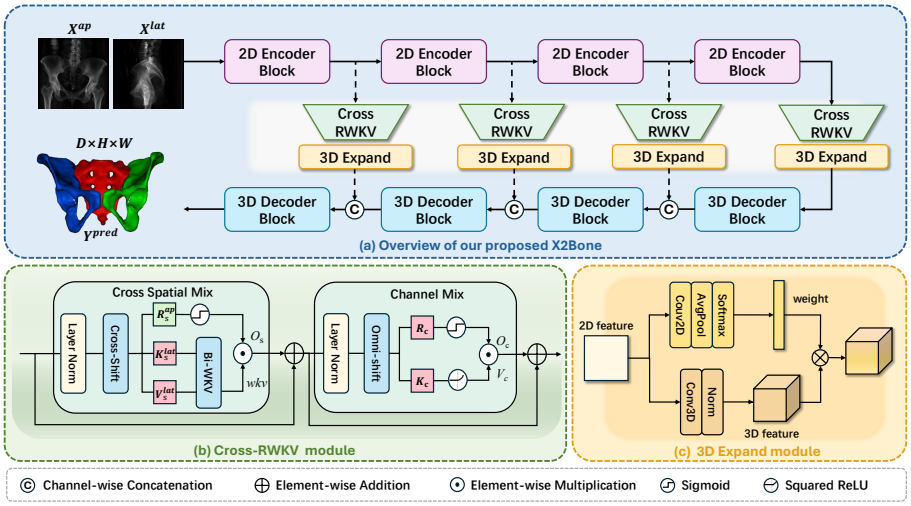

Fig. 1. (a) Overall architecture of the proposed X2Bone framework; (b) the Cross-RWKV module for long-range cross-view feature interaction; (c) the 3D Expand module for lifting fused 2D features into volumetric space.

2. We design a Cross-RWKV interaction mechanism that enables explicit global alignment and fusion of biplanar features, effectively alleviating structural ambiguity under limited projection conditions.
3. We introduce a depth-attention-driven 2D–3D Lifting Module to reduce information loss during the dimensional expansion from 2D projections to 3D representations.

2 Method

2.1 Architecture Overview

We first present an overview of the encoder–decoder architecture, and subsequently provide detailed descriptions of the individual components in the following subsections. Fig. 1(a) illustrates the overall framework of the proposed X2Bone. Given a pair of AP and LAT X-ray images from the same subject, denoted as X^{ap} and X^{lat}, respectively, the two views are processed in parallel by the encoder to perform hierarchical feature extraction. The extracted 2D features are then fed into the *Cross-RWKV* module to enable effective information interaction and fusion across views. Subsequently, the fused 2D features are lifted to volumetric space via a *3D Expand* module, resulting in voxel-wise 3D feature representations. Finally, a 3D decoder progressively upsamples the features to restore the original spatial resolution, while skip connections are employed to fuse multi-level features, producing the predicted 3D pelvic mask $Y^{\mathrm{pred}} \in \mathbb{R}^{D\times H\times W}$.

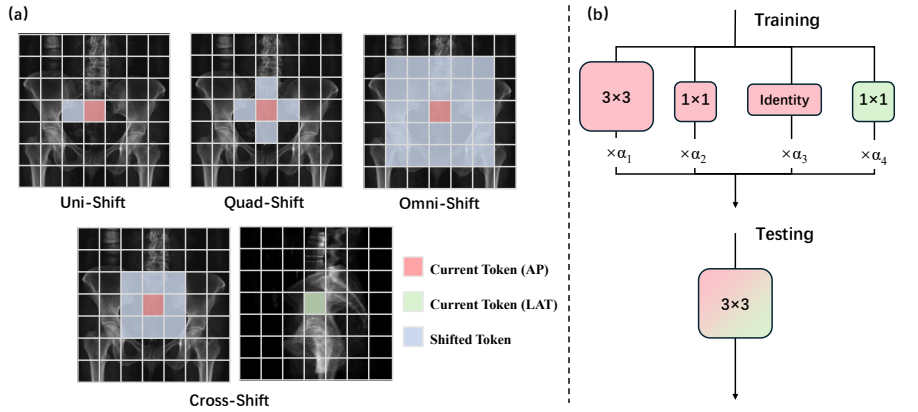

Fig. 2. Illustration of the proposed Cross-shift. (a) Comparison between Uni-Shift[18], Quad-Shift[4], Omni-Shift[22], and our Cross-Shift. (b) Illustration of Cross-Shift with structural re-parameterization.

2.2 Cross-RWKV module

AP and LAT views exhibit strong structural complementarity. However, most existing biplanar 3D reconstruction methods process them as independent branches, which limits the modeling of fine-grained cross-view correlations. To address this limitation, we propose a Cross-RWKV module that explicitly enables long-range cross-view interaction via a linear-complexity bidirectional Receptance Weighted Key–Value (Bi-WKV) mechanism, as illustrated in Fig. 1(b).

Cross spatial mix. Given the encoded feature maps from the AP and LAT branches, we reshape them into token sequences $\mathbf{X}^{\mathrm{ap}}$ and $\mathbf{X}^{\mathrm{lat}} \in \mathbb{R}^{T\times C}$, where T and C denote the sequence length and channel dimension, respectively.

As illustrated in Fig. 2, before reshaping into token sequences, we introduce a Cross-shift operation to exchange and align local spatial context between the two views for cross-view feature aggregation. During training, the module adopts a lightweight re-parameterizable multi-branch structure consisting of a 3×3 convolution, two parallel 1×1 convolutions corresponding to the AP and LAT views, and an identity mapping. Each branch is equipped with a learnable scaling coefficient α to adaptively balance branch contributions during training. At inference time, all branches are analytically merged into a single 3×3 convolution, introducing no additional computational overhead.

Taking the AP (resp. LAT) branch as the target, we generate the receptance gate from the opposite view and compute the key/value from the target view, enabling gated bidirectional cross-view aggregation:

$$\begin{aligned} \tilde{\mathbf{X}}^{\mathrm{ap}},\ \tilde{\mathbf{X}}^{\mathrm{lat}} &= \text{Cross-shift}\left(\mathbf{X}^{\mathrm{ap}}, \mathbf{X}^{\mathrm{lat}}\right), \\ \mathbf{R}^{\mathrm{lat}} = \tilde{\mathbf{X}}^{\mathrm{lat}} W_R, \quad \mathbf{K}^{\mathrm{ap}} &= \tilde{\mathbf{X}}^{\mathrm{ap}} W_K, \quad \mathbf{V}^{\mathrm{ap}} = \tilde{\mathbf{X}}^{\mathrm{ap}} W_V, \\ \mathbf{R}^{\mathrm{ap}} = \tilde{\mathbf{X}}^{\mathrm{ap}} W_R, \quad \mathbf{K}^{\mathrm{lat}} &= \tilde{\mathbf{X}}^{\mathrm{lat}} W_K, \quad \mathbf{V}^{\mathrm{lat}} = \tilde{\mathbf{X}}^{\mathrm{lat}} W_V. \end{aligned} \tag{1}$$

where $W_R, W_K, W_V \in \mathbb{R}^{C\times C}$ are learnable projection matrices. With the above projections, the cross-view spatial aggregation is formulated as

$$\begin{aligned}\mathbf{O}_s^{\text{ap}} &= \sigma(\mathbf{R}^{\text{lat}}) \odot \text{Bi-WKV}(\mathbf{K}^{\text{ap}}, \mathbf{V}^{\text{ap}}),\\ \mathbf{O}_s^{\text{lat}} &= \sigma(\mathbf{R}^{\text{ap}}) \odot \text{Bi-WKV}(\mathbf{K}^{\text{lat}}, \mathbf{V}^{\text{lat}}).\end{aligned} \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid gating function and $\odot$ represents element-wise multiplication. $\mathbf{O}_s^{\text{ap}}$ and $\mathbf{O}_s^{\text{lat}}$ denote the cross-view spatial responses of the AP and LAT branches, respectively. For notation simplicity, we denote the output of the cross spatial mix as $\mathbf{O}_s = (\mathbf{O}_s^{\text{ap}}, \mathbf{O}_s^{\text{lat}})$, and the subsequent channel-mix operations are applied to the two branches separately.

The bidirectional weighted key–value operator aggregates global context across the entire token sequence. For the t-th token, Bi-WKV is defined as

$$\text{Bi-WKV}(K,V)_t = \frac{\sum_{i=0,i\neq t}^{T-1} e^{-(|t-i|-1)/T\cdot w + k_i}\, v_i \;+\; e^{u+k_t}\, v_t}{\sum_{i=0,i\neq t}^{T-1} e^{-(|t-i|-1)/T\cdot w + k_i} \;+\; e^{u+k_t}}. \tag{3}$$

Here, $k_i, v_i \in \mathbb{R}^C$ denote the key and value of the i-th spatial token, $w \in \mathbb{R}^C$ is a learnable channel-wise decay vector controlling distance attenuation, and $u \in \mathbb{R}^C$ is a learnable bias emphasizing the current token. All exponential and multiplication operations are performed channel-wise.

Channel mix. Subsequently, the spatial responses are passed into the channel-mix module to perform feature interaction along the channel dimension. Following the RWKV formulation, the receptance and key are generated by applying layer normalization and Omni-Shift[22] to the input tokens:

$$\mathbf{R}_c = \text{Omni-Shift}(\text{LN}(\mathbf{O}_s))W_{R_c}, \quad \mathbf{K}_c = \text{Omni-Shift}(\text{LN}(\mathbf{O}_s))W_{K_c}, \tag{4}$$

where $W_{R_c}, W_{K_c} \in \mathbb{R}^{C\times C}$ are learnable projection matrices.

The channel-mixed output is computed via a gated channel-wise transformation:

$$\mathbf{O}_c = \big(\sigma(\mathbf{R}_c) \odot \mathbf{V}_c\big)\, W_{O_c}, \quad \text{where } \mathbf{V}_c = \gamma(\mathbf{K}_c)\, W_{V_c}. \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\gamma(\cdot)$ is the Squared ReLU activation, and W_{V_c}, W_{O_c} are learnable projection matrices. Residual connections are employed to facilitate stable optimization in deep networks.

2.3 3D Expand module

To address the dimensional mismatch between 2D planar features and the 3D voxel space, we introduce a 3D Expand module, as depicted in Fig. 1(c). Given the fused 2D feature map $\mathbf{X}_{2d} \in \mathbb{R}^{C\times H\times W}$, the module first predicts a global distribution over the missing spatial dimension by applying a 2D convolution followed by global average pooling and Softmax normalization:

$$\mathbf{w}_d = \text{Softmax}\big(\mathcal{P}_{avg}(\text{Conv}_{2d}(\mathbf{X}_{2d}))\big), \tag{6}$$

where $\mathbf{w}_d \in \mathbb{R}^D$ represents the learned probability weights along the spatial dimension that is unobserved in the current view.

Simultaneously, the 2D features are lifted into an initial volumetric representation by applying a 3D transposed convolution along the missing spatial dimension, yielding $\mathbf{V}_{raw} \in \mathbb{R}^{C\times D\times H\times W}$. Depending on the input view (AP or LAT), a view-specific permutation is applied to align the generated dimension with the spatial axis associated with the predicted probability weights. Finally, the volumetric feature is obtained by modulating the expanded volume along this aligned dimension using the learned weights:

$$\mathbf{V}_{out} = \mathbf{V}_{raw} \odot \mathcal{T}_{view}(\mathbf{w}_d), \tag{7}$$

where $\mathcal{T}_{view}(\cdot)$ denotes the view-dependent alignment operation. This operation is performed for both AP and LAT views, producing two expanded volumetric features $\mathbf{V}_{out}^{ap}$ and $\mathbf{V}_{out}^{lat}$, respectively. The final 3D feature representation is obtained by element-wise summation:

$$\mathbf{V}_{3d} = \mathbf{V}_{out}^{ap} + \mathbf{V}_{out}^{lat}, \tag{8}$$

where $\mathbf{V}_{3d}$ denotes the fused volumetric feature. This mechanism enables the model to compensate for view-missing spatial information and integrate complementary cues from both AP and LAT views, leading to a more accurate reconstruction of the 3D volumetric representation.

2.4 Encoder and Decoder

Both the 2D encoder block and the 3D decoder block consist of two successive convolutional layers, each followed by batch normalization and ReLU activation. Residual connections are employed to fuse the input and output features, enhancing feature representation while facilitating stable gradient propagation. In addition, the 3D decoder concatenates the expanded volumetric features from the 3D Expand module with skip-connected decoder features to preserve fine-grained 3D details during upsampling.

3 Experiments and Results

3.1 Settings

Datasets. Experiments were conducted on the CTPelvic1K dataset [12] to evaluate the proposed method under bi-planar X-ray settings. We reconstruct the left hip bone, right hip bone, and sacrum, and discard cases with incomplete pelvic structures. The resulting dataset contains 779 training and 197 test samples. Bi-planar X-ray images were simulated as digitally reconstructed radiographs (DRRs) using the ASTRA toolbox [21].

Implementation Details. All experiments were conducted on a single NVIDIA RTX 3090 GPU. We trained the model for 100 epochs using Adam[10] with an

Table 1. Quantitative comparison on the CTPelvic1K biplanar X-ray 3D pelvic bone reconstruction task. Standard deviations are reported for average scores; best results are in bold.

Methods	DSC (%) ↑				HD95 (mm) ↓				Params
	L-Hip	R-Hip	Sac.	Avg.	L-Hip	R-Hip	Sac.	Avg.	(M)
UNet [19]	85.99	84.62	83.90	$84.84_{\pm 4.57}$	4.64	4.79	4.48	$4.64_{\pm 2.56}$	19.22
Att-UNet [15]	87.81	86.88	86.28	$86.99_{\pm 3.33}$	4.23	4.17	4.06	$4.15_{\pm 3.27}$	23.62
UNETR [8]	82.67	81.62	82.58	$82.29_{\pm 5.15}$	6.38	6.86	5.09	$6.11_{\pm 4.30}$	124.49
SwinUNETR [7]	85.89	85.38	85.22	$85.50_{\pm 3.84}$	5.17	4.92	4.27	$4.79_{\pm 2.65}$	62.18
BX2S-Net [3]	87.95	86.98	86.28	$87.07_{\pm 4.01}$	3.92	4.24	3.86	$4.01_{\pm 1.86}$	28.99
SdAOF [2]	87.11	87.91	86.60	$87.21_{\pm 3.14}$	4.67	4.04	4.09	$4.27_{\pm 2.72}$	416.43
X2Bone (ours)	**88.68**	**87.96**	**87.23**	$\mathbf{87.96}_{\pm 3.01}$	**3.58**	**3.86**	**3.68**	$\mathbf{3.70}_{\pm 1.26}$	**11.02**

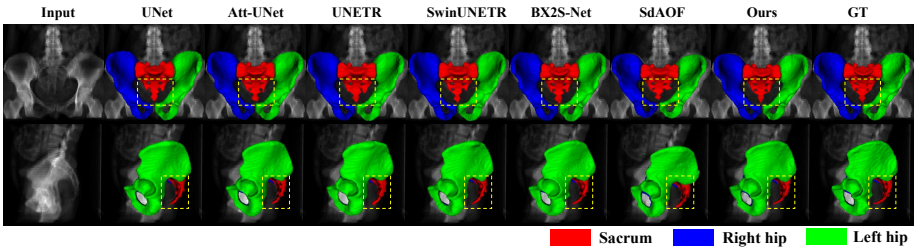

Fig. 3. Qualitative visualization results on the CTPelvic1K dataset. The first and second rows correspond to AP and LAT DRRs, respectively, with reconstructed 3D pelvic structures and ground truth under the same views.

initial learning rate of 1×10^{-4}. Performance was evaluated by Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95). Paired AP and LAT X-rays were resized to 128×128 with matched in-plane spacing, and CT segmentations were resampled to isotropic 128^3 voxels for alignment. Unless otherwise specified, only AP and LAT views were used as input.

3.2 Performance

We compare the proposed X2Bone with a range of recent state-of-the-art methods. Table 1 reports the DSC and HD95 scores on the CTPelvic1k dataset, together with the parameters for each model, providing a comprehensive evaluation of reconstruction accuracy and model complexity. The results show that X2Bone achieves highly competitive overall performance, obtaining the highest average DSC score of 87.96 and the lowest HD95 of 3.70, outperforming most existing approaches. Notably, compared with the latest state-of-the-art method SdAoF, X2Bone uses only 11.02M parameters, demonstrating significantly higher efficiency while maintaining superior reconstruction accuracy.

Figure 3 presents qualitative visualization results, including the input AP and LAT DRRs and the corresponding reconstructed 3D organ shapes. As highlighted by the yellow dashed boxes (e.g., the distal region of the sacrum and its internal structures), most competing methods suffer from evident under-

Table 2. Ablation study of the proposed components on the CTPelvic1K dataset.

Setting	DSC ↑	HD95 ↓	Params.
Baseline	86.16	4.12	**3.68**
w/o Cross-RWKV	87.21	3.89	8.68
w/o 3D Expand	86.78	4.05	6.02
X2Bone (Ours)	**87.96**	**3.70**	11.02

Table 3. Ablation study on the influence of the number of views on the CTPelvic1K dataset.

Num of views	DSC↑	HD95↓
$N = 1$	87.96	3.70
$N = 2$	89.28	3.30
$N = 3$	89.91	3.19
$N = 6$	**90.55**	**3.02**

segmentation. In contrast, the proposed method effectively alleviates topological discontinuities, producing reconstructions that exhibit higher consistency with the ground-truth annotations in both global morphology and local structural details.

3.3 Ablation Studies

Table 2 evaluates the contribution of the Cross-RWKV and 3D Expand modules on the CTPelvic1K dataset. Compared with the baseline model, introducing either module consistently improves performance, with 3D Expand providing a more pronounced gain by achieving a DSC score of 87.21 and reducing HD95 to 3.89, highlighting the importance of explicit 2D-to-3D feature lifting for spatial structure modeling. When both modules are enabled simultaneously, the model performance is further improved, reaching a DSC score of 87.96 and an HD95 of 3.70. These results demonstrate that Cross-RWKV and 3D Expand exhibit strong complementarity in enhancing model performance.

We further investigate the influence of the number of input views on reconstruction performance, as reported in Table 3. To construct multi-view inputs, DRRs are generated by rotating the X-ray source around the target anatomy. Starting from the coronal and sagittal planes of the CT volume, N views are rendered at an angular interval of $90/N$ degrees for each plane. When $N = 1$, this setting corresponds to the standard orthogonal biplanar configuration consisting of one AP view and one LAT view. As the number of views increases, reconstruction accuracy consistently improves, with the DSC score rising from 87.96 to 90.55 and HD95 decreasing from 3.70 to 3.02 when six views are used. This trend indicates that additional views provide complementary geometric information that helps reduce depth ambiguity and anatomical overlap inherent in 2D X-ray projections. Notably, the performance gain gradually saturates as more views are introduced, suggesting that the proposed X2Bone can efficiently aggregate multi-view information without requiring a large number of inputs.

4 Conclusion

In this work, we propose X2Bone, a novel framework for 3D bone reconstruction from biplanar X-ray images. By integrating the Cross-RWKV interaction

mechanism with an efficient 2D–3D feature lifting strategy, X2Bone achieves state-of-the-art reconstruction performance with a lightweight model footprint. Future work will extend the framework to other skeletal structures and focus on building paired real X-ray and CT datasets for validation in clinical scenarios.

Acknowledgments. This work is partly supported by grants from the National Key Research and Develop Program of China (2023YFC2411500), Shenzhen Science and Technology Program (JCYJ20241202124902004), and Basic and Applied Basic Research Foundation of Guangdong (2024A1515010695).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aubert, B., Vazquez, C., Cresson, T., Parent, S., de Guise, J.A.: Toward automated 3d spine reconstruction from biplanar radiographs using cnn for statistical spine model fitting. IEEE transactions on medical imaging **38**(12), 2796–2806 (2019)
2. Chen, J., Lin, Y., Sun, H., Li, X.: Spatial-division augmented occupancy field for bone shape reconstruction from biplanar x-rays. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 668–678. Springer (2024)
3. Chen, Z., Guo, L., Zhang, R., Fang, Z., He, X., Wang, J.: Bx2s-net: Learning to reconstruct 3d spinal structures from bi-planar x-ray images. Computers in biology and medicine **154**, 106615 (2023)
4. Duan, Y., Wang, W., Chen, Z., Zhu, X., Lu, L., Lu, T., Qiao, Y., Li, H., Dai, J., Wang, W.: Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. arXiv preprint arXiv:2403.02308 (2024)
5. Fei, Z., Fan, M., Yu, C., Li, D., Huang, J.: Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. arXiv preprint arXiv:2404.04478 (2024)
6. Frija, G., Blažić, I., Frush, D.P., Hierath, M., Kawooya, M., Donoso-Bach, L., Brkljačić, B.: How to improve access to medical imaging in low-and middle-income countries? EClinicalMedicine **38** (2021)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
8. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
9. Karade, V., Ravi, B.: 3d femur model reconstruction from biplane x-ray images: a novel method based on laplacian surface deformation. International journal of computer assisted radiology and surgery **10**(4), 473–485 (2015)
10. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Liu, J., Li, H., Zeng, B., Wang, H., Kikinis, R., Joskowicz, L., Chen, X.: An end-to-end geometry-based pipeline for automatic preoperative surgical planning of pelvic fracture reduction and fixation. IEEE Transactions on Medical Imaging (2024)

12. Liu, P., Han, H., Du, Y., Zhu, H., Li, Y., Gu, F., Xiao, H., Li, J., Zhao, C., Xiao, L., et al.: Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. International Journal of Computer Assisted Radiology and Surgery **16**(5), 749–756 (2021)
13. Mancino, F., Fontalis, A., Magan, A., Plastow, R., Haddad, F.S.: The value of computed tomography scan in three-dimensional planning and intraoperative navigation in primary total hip arthroplasty. Hip & pelvis **36**(1), 26 (2024)
14. Mitton, D., Landry, C., Veron, S., Skalli, W., Lavaste, F., De Guise, J.A.: 3d reconstruction method from biplanar radiography using non-stereocorresponding points and elastic deformable meshes. Medical and biological engineering and computing **38**(2), 133–139 (2000)
15. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
16. Ou, X., Chen, X., Xu, X., Xie, L., Chen, X., Hong, Z., Bai, H., Liu, X., Chen, Q., Li, L., et al.: Recent development in x-ray imaging technology: Future and challenges. Research (2021)
17. Paxton, N.C., Nightingale, R.C., Woodruff, M.A.: Capturing patient anatomy for designing and manufacturing personalized prostheses. Current Opinion in Biotechnology **73**, 282–289 (2022)
18. Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., et al.: Rwkv: Reinventing rnns for the transformer era. arXiv preprint arXiv:2305.13048 (2023)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
20. Shakya, M., Khanal, B.: Benchmarking encoder-decoder architectures for biplanar x-ray to 3d bone shape reconstruction. Advances in Neural Information Processing Systems **36**, 20469–20481 (2023)
21. Van Aarle, W., Palenstijn, W.J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., De Beenhouwer, J., Joost Batenburg, K., Sijbers, J.: Fast and flexible x-ray tomography using the astra toolbox. Optics express **24**(22), 25129–25147 (2016)
22. Yang, Z., Li, J., Zhang, H., Zhao, D., Wei, B., Xu, Y.: Restore-rwkv: Efficient and effective medical image restoration with rwkv. IEEE Journal of Biomedical and Health Informatics (2025)
23. Ye, H., Tang, F., Zhao, P., Huang, Z., Zhao, D., Bian, M., Zhou, S.K.: U-rwkv: Lightweight medical image segmentation with direction-adaptive rwkv. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 613–623. Springer (2025)