BENCHMARKING ENCODER-DECODER ARCHITECTURES FOR BIPLANAR X-RAY TO 3D BONE SHAPE RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Various deep learning pipelines have been proposed for 3D Bone Shape Reconstruction from Biplanar X-rays. Although these methods individually report excellent results, we do not know how these architecture pipelines compare against each other since they are reported on different anatomy, datasets, and cohort distribution. We benchmark these disparate architectures on equal footing on three different anatomies using public datasets. We describe various benchmarking tasks to simulate real-world clinical scenarios including reconstruction of fractured bones, bones with implants, robustness to population shift, and estimate clinical parameters. We provide an open-source implementation of SOTA architectures, dataset pipelines, and extraction of clinical parameters. Comparing the encoder-decoder architectures with baseline retrieval models, we find that the encoder-decoder methods are able to learn from data and are much better than retrieval baselines. However, the best methods have limited difference on performance, but the domain shift plays an important role in deteriorating the performance of these methods.

1 INTRODUCTION

X-ray is the most common and widely used imaging modality for orthopedics, trauma, and dentistry as it has low radiation, low cost, and is portable. X-ray scanner projects 3D information of the target body into a plane, resulting in a 2D image. This 2D representation is not ideal and sometimes not enough for visualizing the 3D structure that can be important in diagnosis, prognosis, surgery planning and navigation, and medical education. CT scan captures X-ray-like images from several angles covering 360 degrees and reconstructs a single volumetric image, providing detailed 3D structural information of the target anatomy. However, CT scan has relatively high radiation, is costly, and is not even available in many rural health centers across the globe. Hence, there has been a longstanding interest in the scientific community to develop methods that can reconstruct 3D images or structures of interest from few to single X-ray images of various human bones (Hindmarsh, 1973; Suh, 1974; Shiode et al., 2021), teeth (Song et al., 2021), and anatomies of other species (Henzler et al., 2018).

From the early days of the stereo-correspondence point-based approach (Brown et al., 1976; Pearcy, 1985), several methods have been proposed for 3D reconstruction from biplanar radiographs: non-stereo-corresponding point-based, contour-based, statistical shape model (SSM) based, parametric, and hybrid methods (Hosseinian & Arefi, 2015). These methods face challenges in extracting hand-crafted features (landmarks or contours), using deformation priors and atlas, or building SSMs (Mitton et al., 2000; Benameur et al., 2003; Karade & Ravi, 2015; Aubert et al., 2019). Recently, several deep learning-based approaches show promise in providing more accurate results with faster computation time (Melissa et al., 2021). Although methods combining deep learning and anatomical priors via atlas or features such as landmarks or contours are starting to emerge (Chênes & Schmid, 2021; Bayat et al., 2022; Van Houtte et al., 2022), most of the existing deep-learning-based methods predominantly predict shape or image directly using an encoder-decoder-based neural network architecture (Bayat et al., 2020; Kasten et al., 2020; Nakao et al., 2021; Shiode et al., 2021; Almeida et al., 2021).

Although deep learning based 2D-3D reconstruction from biplanar X-ray is an active research topic with several different models proposed (Bayat et al., 2020; Kasten et al., 2020; Nakao et al., 2021; Shiode et al., 2021; Almeida et al., 2021), we still do not know which methods are the best ones for what contexts. We observe four main challenges that need to be addressed for better progress in this field: i) *Lack of a platform* that brings together publicly available dispersed datasets of different anatomies in different standards to a common standard; this makes it difficult for researchers developing new methods to evaluate their methods in a common benchmark. ii) *Lack of reproducible works* due to the use of private datasets, incomplete description of hyperparameters, and closed source code. iii) *Lack of a comprehensive evaluation* required to assess the potential of clinical translation: evaluation on single anatomy which is different for different papers; aggregated results without providing insight on the model's performance across different types of data such as pathology, type of structure (such as different vertebra type), number of samples of each type, image resolution, etc. iv) Limited effort towards *identifying and optimizing clinical parameters* that are of interest: most methods use metrics such as Dice Score or Hausdroff's Distance for measuring reconstruction accuracy, but very limited study on how these metrics actually impact the clinical parameters and decision making. For instance, it is not clear at all if the dice score of 0.8 instead of 0.9 would bring any difference to an actual clinical application where the method is applied.

In this work, we bring together four dispersed and different multi-center publicly available datasets into a common standard, and benchmark various recently proposed encoder-decoder architectures on this dataset. We evaluate the methods across different anatomy, analyze results across multiple metrics segregated by various factors of variation that are of clinical interest, and assess their ability to estimate important clinical parameters.

Procedure Reference→	Previous Approaches			
	UNet Kasten et al.	Transvert Bayat et al.	TL-Embedding Shiode et al.	Encoder-Decoder Chen & Fang
Anatomy of Interest	Knee	Vertebra	Wrist	Vertebra
Input views	AP & LAT	AP & LAT	PA	AP & LAT
Input DRR Res(mm)	1.0	1.5	0.4	1.5
Input Size	128^2	64^2	500×625	64^2
Training Samples	188	$\sim$10k	147	90
Test Samples	20	$\sim$2k	26	10
Test Sample Res(mm)	1.0	1.0-3.0	0.625,1.25	1.5
Supervised Loss	weighted-CE	L1	CE	L2
Adversarial Loss	✗	✓	✗	✗
Reprojection Loss	✓	✗	✗	✗
AP/LAT View-Fusion	Input-level	Feature-level	AP view only	Feature-level
Data Augmentation	$\pm 5^\circ$		$\pm 30^\circ$	$\pm 90^\circ$
Surface Error(mm)				
↳ avg	1.075 - 1.709	✗	1.05 - 1.45	✗
↳ max	✗	5.11	✗	✗
Dice Score	84.8 - 94.5	✗	✗	✗
↳avg. across classes	90.7	95.5	✗	74.0

Table 1: Ingredients and hyper-parameters used for biplanar x-ray to 3D Bone shape Reconstruction used in prior works: We see that the original works were validated on disparate dataset size, resolution, dataset cohort, different bone anatomies for different papers etc. resulting in non-comparable reported performance. Many of these disparate architectures report good surface reconstruction($< 2mm$ average surface reconstruction error)

Contributions *Provide comprehensive and standardized framework for datasets access, preprocessing, baseline comparison, and a set of metrics:* we provide an open-source implementation for accessing four different publicly available datasets, bring them into a common standard and preprocessing pipeline, provide reference implementation of SOTA architecture, analysis scripts for extracting clinically relevant parameters from reconstructed bone shape.

Benchmark Encoder-Decoder based Architectures: we benchmark encoder-decoder based end-to-end architectures which were proposed specifically for x-ray to 3D bone shape reconstruction,

namely Kasten et al. (2020); Bayat et al. (2020); Shiode et al. (2021), on three anatomies (vertebra, hip and knee) using public datasets.

Provide major anatomical and pathological categories for disaggregated reporting of the results, and highlight the limitations of the commonly followed approach of reporting aggregated results: specifically, we report disaggregated results on vertebra sub-types (cervical, thoracic and lumbar), vertebra sub-structure (vertebral body, vertebral arch and transverse processes).

Robustness to Domain shifts We describe various benchmarking tasks to perform an external validation of the baselines in a realistic clinical context such as when the x-ray contains fractured bones or implants, or when the model is evaluated on an entirely different population cohort.

2 Related Work

Recent studies have highlighted the issues in deep learning research with too many architectures without a clear understanding of which one is better or disguising speculation as explanations (Lipton & Steinhardt, 2019), lack of reproducibility (Bouthillier et al., 2019), a gap in improvements in benchmarks vs actual applications such as in clinical translations (Varoquaux & Cheplygina, 2022). These studies have different scopes, such as the broad scope of machine learning in general but focusing on reproducibility (Bouthillier et al., 2019) or field-specific contexts such as in health care covering broad issues (Varoquaux & Cheplygina, 2022). Such recent works covering a broad field provide guidelines and emphasize important topics that need attention from the scientific community. Nevertheless, specific applications can have peculiar needs and issues that require further investigation and attention. For example, the COVID-19 pandemic led to a very large number of papers on deep learning-based COVID-19 detection in a short period of time, but almost all these methods were clinically not useful and had study design flaws and bias (Roberts et al., 2021; Hasan et al., 2022). In this work, we focus on the task of 3D reconstruction of human bones and joints from a pair of 2D X-ray images.

3 Four Datasets from Three Different Bones

Digitally Reconstructed Radiographs (DRRs) as input images The machine learning (ML) task being studied in this work can be described as when given a human subject's pair of Anterio-Posterior (AP) and Lateral X-ray images of a target bone, reconstructing the 3D structure of the target bone. For an ideal training and evaluation of a model for this task, ground truth 3D structure would be the expert manual segmentation from a CT scan of the patient whose AP and lateral images are taken as input by the model. However, to our knowledge, there are no publicly available paired images of real X-ray scans and CT scans of the same patients. Since most methods construct AP and lateral Digitally Reconstructed Radiographs (DRRs) from CT scan image and use it as input to the model for training and evaluation, we use the same approach in this work. In this work, input to all ML models are DRRs generated from 3D scans using Siddon-Jacobs Raytracing Algorithm (Wu, 2010).

DRRs from CT scans We create two datasets for which input DRR images are generated from CT scans, placing the CT volume midway between the simulated x-ray source and the image plane which is separated by a source-to-image distance of 200cm. In order to get cleaner delineation of bones with fewer attenuation from soft tissues, only those voxels whose Hounsfield Unit(HU) > 0 are taken into consideration to calculate ray attenuation.

DRRs from ground truth segmentation masks: In addition to input DRRs generated from CT scans which are closer to real x-ray scans, we create two additional datasets where DRRs are generated from segmentation masks, named DRRfromMask. These DRRfromMask provide simpler projections of the target 3D shape, which is useful to study how different models perform when the reconstruction task is simpler without any confounding non-bone structures in the input images.

3.1 VERSE2019-SPINE-DRR

VerSe2019 (Sekuboyina et al., 2021) contains 160 CT scans with two types of annotations: manual segmentation of individual vertebra, and the centroid location provided as metadata. The vertebra with foreign materials such as cement and screws do not have manual annotations, which when removed results in 1722 vertebra from the 160 CT scans.

VerSe2019-Spine-DRR Train-Val-Test Split The 160 CT scans are randomly divided into four groups of 40 images each. In a standard, 4-fold cross-validation setup, this would result in four different groups of 120 scans for training and 40 scans for validation. However, we use four groups of 40 CT scans for the test. For internal validation, the 120 scans are further split into 110 training and the remaining 10 for hyperparameter tuning and early stopping. Since the input to the model is an individual vertebra instead of the whole CT scan, the vertebra in all the images that contain foreign materials is excluded during training and testing by using the metadata provided in the VerSe2019 dataset.

For training, there are on average 1451 vertebra from the four groups of 110 images in the training set, and in the test set, there are 271 vertebra from the four groups of 40 CT scans.

3.2 VERSE2019-SPINE-DRRFROMMASK

This dataset is exactly the same as VerSe2019-Spine-DRR described in the previous section but with one difference: the DRR images are generated from the ground truth segmentation masks instead of the CT scan images.

3.3 CTPELVIC1K-DRR

CTPelvic1k (Liu et al., 2021) consists of 1106 pelvic CT scans with manual segmentation of left and right pelvic bones, sacrum, and vertebra near the sacrum. The dataset has seven subsets (number of scans in parenthesis): *Adbomen*(35), *Colonog* (714), *Msd-T10* (155), *Kits* (44), *Cervix* (41), *Clinic* (103), and *Clinic-metal* (14). *Clinic* consists of scans with fractured bones, and *Clinic-metal* has images having foreign bodies such as implants, screws, and rods. *Msd-T10* was unusable as it had highly anisotropic voxels which resulted in unrealistic and pixelated DRR.

CTPelvic1k-DRR Train-Val-Test Split The largest subset *Colonog* containing 714 images is randomly split into 80:20, resulting in training and a validation set of 535 and 179 images respectively. There was an uneven number of vertebra segmented, hence we changed the vertebra segmentation into the background label. Similarly, the pelvic bone and sacrum labels were merged to produce a binary mask for the hip excluding the vertebra. The other subsets are used as test images and to assess various models' robustness to domain shift.

3.4 OAI-ZIB-KNEE-DRRFROMMASK

This dataset consists of 507 knee MRI images with manual segmentation, where the manual segmentation is provided by Ambellan et al. (2019) for the 507 MRIs selected from a larger database (Peterfy et al., 2008).

OAI-ZIB-Knee-DRRfromMask Train-Val-Test Split The total of 507 images is split into 380 training and 179 validation images.

4 ENCODER-DECODER AND BASELINE MODELS

Encoder-Decoder-based architecture is the most popular deep neural network used by state-of-the-art X-ray to 3D reconstruction methods. We implement and evaluate four state-of-the-art architectures that use different approaches, and use different datasets for evaluating their respective model's performance: Kasten et al. (2020), Bayat et al. (2020), Shiode et al. (2021) and Chen & Fang (2019). These architectures introduce inductive bias in different ways that can be characterized by: i) whether input orthogonality is respected, and ii) at what stage of the encoder-decoder pipeline the

two input views are fused. Kasten et al. (2020) fuses the two views at the input stage respecting the orthogonality of the inputs to obtain a 3D-like input allowing the use of popular off-the-shelf 3D-to-3D Encoder-Decoder architectures such as U-Net. Bayat et al. (2020), on the other hand, fuses the two views at a later stage where the encoder has generated high-level features. Chen & Fang (2019) applies the Encoder-Decoder framework fusing the two views by concatenating the low-dimensional 1-D embedding vector. Shiode et al. (2021) uses an auto-encoder (T-Network) to learn the distribution of the 3D segmentation in a low-dimensional manifold (Girdhar et al., 2016) and then projects the X-ray into the same learned manifold, using a separate 2D image encoder (L-Network). At test-time, the L-Network and the decoder part of the T-Network are used to obtain 3D reconstruction from a single X-ray.

Comparison to Retrieval-based Baselines Tatarchenko et al. (2019); Liao et al. (2021) show that simple image classification and retrieval-based baselines outperformed CNN architectures on object reconstruction task in certain scenarios. In order to evaluate how much the encoder architecture help in X-ray to 3D reconstruction task, we implement two baseline methods: Nearest Neighbors and Oracle.

Nearest Neighbour: For a given pair of input X-ray images, it returns the 3D segmentation from the training set of the sample which has the most similar X-ray pair image to the input pair. The similarity is measured as the average sum of the squared differences between corresponding X-ray views.

Oracle Following Henzler et al. (2018), for a given input pair of X-ray images, we use its 3D ground truth shape and return the closest shape in the training set. Although this is impractical for inference in a realistic setting where the ground truth of the input X-ray pair is unknown, it is useful in an evaluation setup where the ground truth of the test set is known and provides an upper threshold on how well retrieval-based methods may work.

5 Experimental Setup

5.1 Preprocessing

The CT-Segmentation pair is cropped to the Region-of-Interest by using a ground-truth bones segmentation bounding box. They are then resampled to the $1mm^3$ voxel dimension. The obtained volume is padded to the nearest power of 2 (128^3 for vertebra and 256^3 for hip bones) and oriented to common LPS orientation. The intensity values of generated DRR are scaled to lie within the range [0,1].

5.2 Hyperparameters

We manually tune the architectural hyperparameters (number of layers, kernel size, number of feature maps, etc.) taking reference from prior work as the starting point. We use Adam Optimizer with an initial learning rate of 2e-3 (chosen from grid search). We keep the exponential moving average of the validation score(val_{MA}) and terminate training if val_{MA} does not improve within the last 20 epochs or within 100 epochs, whichever is the earliest. A dropout rate of 0.1 is chosen for all convolutional layers to prevent overfitting, and PReLU activation is used.

5.3 Implementation Details

For most of the replicated methods, we adhere to the same architectural details as in the original work, except when modifications resulted in improvement. For example, Kasten et al. (2020) uses regular UNet whereas we use Residual UNet(Zhang et al., 2018) improving performance. Similarly, we use combined Dice and Cross-Entropy loss for training all the models resulting in faster convergence. For Bayat et al. (2020), we added additional residual blocks (6 instead of 4) keeping other details intact. For Shiode et al. (2021), exact architectural details were not reported in the original work forcing us to use reasonable choices taking the 11 GB GPU memory budget available.

Figure 1: From Left to Right: i) and ii) Biplanar X-ray images iii) Ground truth 3D Bone Shape iv,v.vi and vii) reconstructed bone shape from the baseline encoder-decoder architectures, viii and ix) reconstructed bone shape from the retrieval-based baselines.

5.4 Evaluation Metrics

We report Dice Metric, Hausdorff Distance(95th percentile), Average Surface Distance(ASD), and Normalized Surface Distance(NSD). These metrics measure shape overlap, maximum and average distance between two surfaces and variability tolerant surface similarity.

Although image metrics such as Dice and HD provide measures on how well the reconstruction is done, it does not necessarily provide a measure for whether it is satisfactory for clinical decision making or not. Estimating clinically meaningful parameters from the predicted reconstruction and comparing it with ground truth clinical parameters can provide a more direct measure for clinicians. For instance, Lerchl et al. (2022) use segmented spine to obtain a musculoskeletal model to estimate forces acting on the vertebra, and use it to identify vertebra landmarks.

Keeping the evaluation of an exhaustive list of clinical parameters for all the anatomy as future work, here we estimate the following two important clinical parameters for pelvic region, Intercristal Distance(ID) and Transverse Diameter of Pelvic Inlet(TDUP), to highlight the importance of such experiments.

6 Results

6.1 Baseline Results on Four Datasets

Table 2 presents the performance of four deep learning models and two retrieval baseline models on four different datasets, resulting on 24 different models that are evaluated using four different metrics. The results show that encoder-decoder models perform better when DRRs are generated from masks rather than from CT scan images. DRRfromMask datasets incorporate only the shape variability and removes the surrounding tissue structure and scanner artifacts. Such simplication seems to aid the networks to reconstruct 3D shapes. Among the two DRRfromMask datasets, the performance of the models are much better in knee dataset compared to the spine dataset despite the fact that the spine dataset has larger number of training set. Spine has more complex and fine detailed structures compared to knee. Thus, more complex and variable anatomy seems to be more difficult to accurately predict compare to simpler ones.

Although the top two models, implemented from Kasten et al. (2020) and Bayat et al. (2020) provide very close results for a simpler task (when using DRRfromMask datasets), the performances on more challenging dataset seems to favor Kasten et al. (2020) whose improvement over Bayat et al. (2020) grows by a substantial margin going from vertebra to hip.

6.2 Disaggregated results on Verse2019-Spine-DRR

Table 3 reports results of various models on Verse2019-Spine-DRR for test images grouped using different classification approaches: vertebra type, compression fracture type, and compression fracture severity. It is interesting to see that despite the fact that Wedge and Concave vertebra, or Mild and Moderate fractures have smaller number of samples (approx. 1000 vs 100), their performance is similar or even better than the normal vertebra. However, the Crush and Severe fractures have much lower performance. The crush fracture results in variation in texture and severe fracture has larger change in appearance compared to other fractures. Such a distinct variation might have played an important role together with their numbers being very low, around 30.

Dataset (#train/#test) (vol size)	Method Reference	#Param	Dice(%)↑	HD95(mm)↓	ASD(mm)↓	NSD@1mm↑
VerSe19-Spine-DRR (1451/271) 128 X 128 X 128	Kasten et al.	1.20M	**84.08** ± 0.23	3.86 ± 0.25	**1.03** ± 0.05	**0.74** ± 0.02
	Bayat et al.	1.91M	83.15 ± 0.42	**3.68** ± 0.19	1.11 ± 0.02	0.72 ± 0.01
	Shiode et al.	1.70M	77.99 ± 0.16	5.43 ± 0.90	1.76 ± 0.18	0.61 ± 0.01
	Chen & Fang	1.60M	75.66 ± 1.11	5.96 ± 0.25	1.66 ± 0.13	0.54 ± 0.01
	Nearest Neighbor	N/A	69.95 ± 1.31	6.91 ± 0.83	2.19 ± 0.30	0.49 ± 0.02
	Oracle	N/A	79.15 ± 0.67	4.33 ± 0.19	1.44 ± 0.05	0.64 ± 0.01
CTPelvic1K-DRR (535/179) 256 X 256 X 256	Kasten et al.	2.97M	**82.52**	**7.66**	**2.13**	**0.53**
	Bayat et al.	2.95M	75.94	10.17	2.45	0.34
	Shiode et al.	9.64M	52.86	29.69	3.70	0.13
	Chen & Fang	4.25M	71.40	11.42	2.97	0.28
	Nearest Neighbor	N/A	57.37	13.00	4.39	0.27
	Oracle	N/A	64.62	10.60	3.57	0.32
OAI-ZIB-Knee-DRRfromMask (380/179) 128 X 128 X 128	Kasten et al.	1.20M	**97.73** ± 0.04	**1.18** ± 0.02	**0.41** ± 0.00	**0.973** ± 0.00
	Bayat et al.	1.91M	97.56 ± 0.03	1.24 ± 0.01	0.44 ± 0.00	0.968 ± 0.00
	Shiode et al.	1.70M	94.21± 0.04	2.63± 0.01	0.87± 0.00	0.753± 0.00
	Chen & Fang	1.60M	93.68 ± 0.18	2.95 ± 0.07	1.03 ± 0.03	0.728 ± 0.01
	Nearest Neighbor	N/A	88.96 ± 0.15	4.88 ± 0.04	1.80 ± 0.02	0.509 ± 0.01
	Oracle	N/A	89.23 ± 0.14	4.76 ± 0.06	1.76 ± 0.02	0.520 ± 0.01
VerSe19-Spine-DRRfromMask (1471/271) 128 X 128 X 128	Kasten et al.	1.20M	87.83 ± 0.22	2.67 ± 0.20	0.77 ± 0.04	0.82 ± 0.02
	Bayat et al.	1.91M	**88.20** ± 0.13	**2.44** ± 0.12	**0.73** ± 0.04	**0.83** ± 0.01
	Shiode et al.	1.70M	84.25 ± 0.25	3.17 ± 0.17	1.06 ± 0.07	0.74 ± 0.01
	Chen & Fang	1.60M	83.32 ± 0.21	3.91 ± 0.29	1.01 ± 0.04	0.69 ± 0.02
	Nearest Neighbor	N/A	77.66 ± 0.78	4.67 ± 0.22	1.52 ± 0.07	0.62 ± 0.01
	Oracle	N/A	79.15 ± 0.67	4.33 ± 0.19	1.44 ± 0.05	0.64 ± 0.01

Table 2: Evaluation of four encoder-decoder models and two retrieval baseline methods on four different datasets using four different evaluation metrics. The models perform better on DRRs generated from mask compared to the ones generated from CT scans. Similarly, the models seem to have higher accuracy when the anatomy is simpler (example: knee vs vertebra).

6.3 Domain Adaptation and Generalization Ability on CTPelvic1K-DRR

Medical images come with different types of domain shifts. Table 4 shows the results of model predictions on various groups of test images which have specific types of domain shift compared to the training set of *Colgone* used from CTPelvic1K-DRR. The trained models here are the same as in the baseline models, but the test sets are now different. The different subgroups in Table 4 represent different types of domain shifts. *Clinic* subset of CTPelvic1K contains images with hip fracture; *Clinic-metal* contains images with foreign implants such as cements, bone implants, screws, and rods; and the remaining subgroups come from different anatomy or geographic locations with different scanners. We observe that there is substantial decrease in performance for all the new subsets.

6.4 Clinical Parameters

We extract two clinically relevant parameters from the reconstructed pelvic bones, namely Intercristal Diameter (ID) and Transverse Diameter of Pelvic Inlet (TDUP). Abnormal Intercristal Diameter is known to be associated with increased risk of breast cancer and ovarian cancer(Barker et al., 2012). Also, TDUP has been shown to correlate with incidence of female pelvic floor dysfunction (FPFD)Siccardi & Bordoni (2019).

As shown in Table 6.4 the error sensitivity of these two parameters have disparate difference with the generic overlap- and surface-error based scores. .

7 Discussion and Conclusion

Lack of publicly available source codes, evaluation on partially open public datasets, and lack of using a common datasets and anatomy for comparing across different 2D-3D reconstruction of bones from X-ray image pairs have made it difficult to assess the existing state-of-the-art models. We have evaluated four state-of-the-art encoder-decoder architectures on four different datasets created from three publicly available datasets of different anatomies. Since the ultimate aim of 2D-3D reconstruction from a pair of X-ray is to be useful for various clinical applications, we presented some important metrics and experimental setups that are important before the models can be used in clinical settings. We see that the domain shift can dramatically reduce the performance of models.

Arch↓	Dice(%)↑			HD95(mm)↓			ASD(mm)↓		
Type→	Cervical	Thoracic	Lumbar	Cervical	Thoracic	Lumbar	Cervical	Thoracic	Lumbar
Count→	177	618	393	177	618	393	177	618	393
Kasten	**75.49**	**84.71**	**85.96**	**4.55**	3.52	4.20	**1.25**	**0.97**	**1.02**
Bayat	73.79	83.89	85.14	4.58	**3.34**	**3.86**	1.36	1.05	1.11
Shiode	66.99	78.84	80.74	8.08	4.77	4.75	2.40	1.59	1.58
Chen	65.00	76.33	78.80	6.42	5.71	6.08	1.87	1.62	1.59
NN	56.60	68.17	73.31	8.67	7.17	6.43	2.50	2.26	2.08
Oracle	70.80	79.10	80.15	4.40	4.12	4.59	1.56	1.37	1.50

Arch↓	Dice(%)↑				HD95(mm)↓				ASD(mm)↓			
Type→	Normal	Wedge	Concave	Crush	Normal	Wedge	Concave	Crush	Normal	Wedge	Concave	Crush
Count→	997	71	95	25	997	71	95	25	997	71	95	25
Kasten	**84.16**	**84.86**	**84.82**	75.92	3.79	3.85	4.16	6.18	**1.00**	**1.04**	**1.04**	1.74
Bayat	83.17	83.94	84.46	**76.31**	**3.63**	**3.60**	**3.62**	**5.33**	1.09	1.11	1.08	**1.69**
Shiode	78.16	79.10	79.14	70.87	5.10	4.57	4.77	6.79	1.67	1.64	1.58	2.26
Chen	75.51	77.67	77.63	69.43	5.95	5.29	5.86	7.55	1.65	1.58	1.58	2.24
NN	70.29	68.50	71.49	64.31	6.84	7.28	6.66	7.11	2.15	2.36	2.27	2.35
Oracle	79.57	78.29	79.07	72.35	4.26	4.62	4.52	5.04	1.40	1.51	1.56	1.84

Arch↓	Dice(%)↑				HD95(mm)↓				ASD(mm)↓			
Type→	Normal	Mild	Moderate	Severe	Normal	Mild	Moderate	Severe	Normal	Mild	Moderate	Severe
Count→	997	96	61	34	997	96	61	34	997	96	61	34
Kasten	**84.16**	**85.67**	**84.08**	77.21	3.79	3.80	4.46	5.81	**1.00**	**0.97**	**1.17**	1.52
Bayat	83.17	85.22	82.94	**77.98**	**3.63**	**3.39**	**4.10**	**4.86**	1.09	1.04	1.24	**1.43**
Shiode	78.16	80.39	77.88	71.80	5.10	4.40	5.09	6.43	1.67	1.54	1.75	2.06
Chen	75.51	78.57	76.85	70.29	5.95	5.40	5.98	7.17	1.65	1.57	1.65	2.04
NN	70.29	71.36	71.85	62.17	6.84	6.81	6.56	7.62	2.15	2.20	2.12	2.57
Oracle	79.57	80.57	78.69	71.42	4.26	4.07	4.75	5.56	1.40	1.43	1.63	1.89

Arch↓	Dice(%)↑				HD95(mm)↓				ASD(mm)↓		
Part→	Overall	Body	Arch	Process	Overall	Body	Arch	Process	Overall	Body	Process
Kasten	84.08	86.40	42.88	51.09	3.86	2.99	4.35	13.33	1.03	0.95	4.47
Bayat	83.15	86.14	51.43	52.36	3.68	2.90	4.58	11.88	1.11	0.97	3.91
Shiode	77.99	83.78	38.67	34.73	5.43	3.33	4.61	15.67	1.76	1.15	5.46
Chen	75.66	81.55	31.46	40.67	5.96	3.81	5.53	15.05	1.66	1.34	5.64

Table 3: **Disaggregated metrics on VerSe19-Spine-DRR**: Average performance of various models on different groups of images classified using three different types of categories. Three types of vertebra, for types of compression fractures and four severity levels. Crush type and Severe fractures have much lower performance compared to others but they have only slightly less number of training samples compared to Wedge and Concave or Mild and Moderate.

Although we have covered major architectures, the natural extension of this work will cover more architectures such as transformer based and registration based methods. Similarly, we demonstrated the evaluation of the models on two clinical parameters of hip anatomy. However, there are several clinical parameters for different anatomies on which the models can be evaluated. Moreover, prospective studies where the the impact of model's reconstruction on clinical decision making such as therapy planning could be simulated to study the actual efficacy of these models compared to the CT scan. Finally, we evaluated the method using DRRs instead of realistic X-rays. Although it is difficult to get paired X-rays and CT scans, some works use GANs to translate DRRs into realistic X-rays and then assess the reconstruction ability of the models. However, this evaluation then is dependent on GAN model's ability to produce realistic images. Thus, in future, studies that incorporate paired images such as from patients who first undergo X-ray as first line of check up then followed by CT scans for more detailed checkups would be valuable. These will be explored in future work.

REFERENCES

Diogo F Almeida, Patricio Astudillo, and Dirk Vandermeulen. Three-dimensional image volumes from two-dimensional digitally reconstructed radiographs: A deep learning approach in lower limb ct scans. *Medical Physics*, 48(5):2448–2457, 2021. 1, 2

Felix Ambellan, Alexander Tack, Moritz Ehlke, and Stefan Zachow. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. *Medical image analysis*, 52:109–118, 2019. 4

CTPelvic1k-DRR Sub-Dataset	Method Reference	Obtained Metrics				Reduction in performance against CTPelvic1K-DRR			
		Dice(%)↑	HD95(mm)↓	ASD(mm)↓	NSD@1mm↑	Dice(%)↑	HD95(mm)↓	ASD(mm)↓	NSD@1mm↑
CLINIC	Kasten et al.	77.02±5.85	12.45±12.90	3.07±2.25	0.44±0.07	5.50	4.79	0.94	0.09
	Bayat et al.	64.00±8.22	13.20±4.37	3.82±1.22	0.24±0.06	11.95	3.03	1.37	0.1
	Shiode et al.	11.79±2.61	59.08±3.68	20.33±2.20	0.04±0.01	0.81	3.28	1.91	0.01
	Chen & Fang	62.48±9.64	13.77± 5.61	4.01±1.91	0.23±0.05	8.92	2.35	1.04	0.05
	Nearest Neighbor	52.53±8.53	14.89±4.42	4.90±1.45	0.25±0.05	4.84	1.89	0.51	0.02
	Oracle	58.61±6.54	13.39±4.04	4.37±1.12	0.29±0.04	6.01	2.79	0.8	0.03
CLINIC-METAL	Kasten et al.	69.53±9.33	29.94±25.30	7.36±8.06	0.36±0.08	14.55	27.08	6.33	0.37
	Bayat et al.	65.68±7.42	12.88±2.80	3.72±1.09	0.24±0.05	10.26	2.71	1.27	0.1
	Shiode et al.	15.38±6.16	56.98±6.67	17.22±3.93	0.04±0.02	2.78	5.38	5.02	0.01
	Chen & Fang	48.13±15.17	23.03±13.70	6.34±4.03	0.15±0.05	23.27	11.61	3.37	0.13
	Nearest Neighbor	41.60±10.80	20.86±7.46	7.43±3.15	0.19±0.04	15.77	7.86	3.04	0.08
	Oracle	60.45±9.94	12.96±5.88	4.45±2.09	0.28±0.06	4.17	2.36	0.88	3.04
ABDOMEN	Kasten et al.	71.52±12.28	26.82±21.71	5.92±4.85	0.36±0.10	11	-19.16	-3.79	0.17
	Bayat et al.	69.33±8.43	14.02±6.94	3.38±1.46	0.27±0.06	6.61	3.85	0.93	0.07
	Shiode et al.	12.87±1.56	61.48±3.94	20.82±1.76	0.04±0.01	0.27	0.88	1.42	0.01
	Chen & Fang	65.02±8.93	14.64±4.82	3.34±1.33	0.23±0.05	6.38	3.22	0.37	0.05
	Nearest Neighbor	56.68±9.00	13.31±3.82	4.24±1.13	0.26±0.05	0.69	0.31	0.15	0.01
	Oracle	62.94±5.95	11.23±2.85	3.56±0.65	0.30±0.04	1.68	0.63	0.01	0.02
CERVIX	Kasten et al.	80.00±6.76	12.04±16.01	3.16±4.90	0.48±0.08	2.52	4.38	1.03	0.05
	Bayat et al.	74.36±5.39	10.72±3.16	2.50±0.79	0.32±0.06	1.58	0.55	0.05	0.02
	Shiode et al.	13.28±3.19	62.62±3.80	21.99±1.65	0.04±0.01	0.68	0.26	0.25	0.01
	Chen & Fang	70.54±5.59	12.13±4.57	2.92±1.13	0.27±0.04	0.86	0.71	0.05	0.01
	Nearest Neighbor	56.28±5.29	13.77±3.19	4.69±0.69	0.27±0.04	1.91	0.77	0.3	0
	Oracle	63.06±3.58	11.15±2.78	3.68±0.85	0.32±0.03	1.56	0.55	0.11	0
KITS19	Kasten et al.	64.63±8.37	31.82±20.29	7.48±7.03	0.30±0.04	17.89	24.16	5.35	0.23
	Bayat et al.	64.41±12.95	21.48±20.82	5.41±5.33	0.24±0.07	11.53	11.31	2.96	0.1
	Shiode et al.	12.90±1.56	61.13±4.33	22.63±2.82	0.04±0.01	0.3	1.23	0.23	0.01
	Chen & Fang	48.32±19.66	26.71±15.30	8.45±6.48	0.17±0.07	23.08	25.29	5.48	0.11
	Nearest Neighbor	51.90±19.82	20.40±19.09	7.87±9.31	0.24±0.09	5.47	7.4	3.48	0.03
	Oracle	55.20±14.91	18.31±12.11	5.97±3.92	0.27±0.07	9.42	7.71	2.4	0.05

Table 4: New test subsets with different types of domain shifts in CTPelvic1K. With domain shift, there is substantial decrease in the performance.

Method	Anatomy	Clinical Parameters	
		ID Error(mm)	TDUP Error(mm)
Kasten	Hip	0.50±2.01	3.59±11.62
Bayat		0.76±2.11	5.62±11.96
Shiode		24.15±5.34	10.35±12.66
Chen		0.92±3.34	8.61±12.84

Table 5: Clinical Parameters extracted from 3D Reconstructed Pelvic bone Shape: Intercristal Distance(ID) and Transverse Diameter of Pelvic Inlet(TDUP). ID Error(TDUP Error) represents the difference of ID(TDUP) of the predicted bones shape with that of the ground truth bone shape. We find that different clinical parameters vary in their sensitivity to global reconstruction error as shown by difference in absolute error for the two clinical parameters.

Benjamin Aubert, Carlos Vazquez, Thierry Cresson, Stefan Parent, and Jacques A de Guise. 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. 1

David JP Barker, Clive Osmond, Kent L Thornburg, Eero Kajantie, and Johan G Eriksson. A possible link between the pubertal growth of girls and prostate cancer in their sons. *American Journal of Human Biology*, 24(4):406–410, 2012. 7

Amirhossein Bayat, Anjany Sekuboyina, Johannes C Paetzold, Christian Payer, Darko Stern, Martin Urschler, Jan S Kirschke, and Bjoern H Menze. Inferring the 3d standing spine posture from 2d radiographs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 775–784. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 9

Amirhossein Bayat, Danielle F Pace, Anjany Sekuboyina, Christian Payer, Darko Stern, Martin Urschler, Jan S Kirschke, and Bjoern H Menze. Anatomy-aware inference of the 3d standing spine posture from 2d radiographs. *Tomography*, 8(1):479–496, 2022. 1

Said Benameur, Max Mignotte, Stefan Parent, Hubert Labelle, Wafa Skalli, and Jacques de Guise. 3d/2d registration and segmentation of scoliotic vertebrae using statistical models. *Computerized Medical Imaging and Graphics*, 27(5):321–337, 2003. 1

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pp. 725–734. PMLR, 2019. 3

Richard H Brown, Albert H Burstein, Clyde L Nash, and Charles C Schock. Spinal analysis using a three-dimensional radiographic technique. *Journal of biomechanics*, 9(6):355–IN1, 1976. 1

Chih-Chia Chen and Yu-Hua Fang. Using bi-planar x-ray images to reconstruct the spine structure by the convolution neural network. In *International Conference on Biomedical and Health Informatics*, pp. 80–85. Springer, 2019. 2, 4, 5, 7, 9

Christophe Chênes and Jérôme Schmid. Revisiting contour-driven and knowledge-based deformable models: Application to 2d-3d proximal femur reconstruction from x-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 451–460. Springer, 2021. 1

Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pp. 484–499. Springer, 2016. 5

Md Kamrul Hasan, Md Ashraful Alam, Lavsen Dahal, Shidhartho Roy, Sifat Redwan Wahid, Md Toufick E Elahi, Robert Martí, and Bishesh Khanal. Challenges of deep learning methods for covid-19 detection using public datasets. *Informatics in Medicine Unlocked*, 30:100945, 2022. 3

Phlipp Henzler, Volker Rasche, Timo Ropinski, and Tobias Ritschel. Single-image tomography: 3d volumes from 2d cranial x-rays. *Computer Graphics Forum*, 37(2):377–388, 2018. doi: https://doi.org/10.1111/cgf.13369. 1, 5

James Hindmarsh. *Roentgen stereophotogrammetry for evaluating the effect of scoliosis treatment*. PhD thesis, Verlag nicht ermittelbar, 1973. 1

S Hosseinian and H Arefi. 3d reconstruction from multi-view medical x-ray images–review and evaluation of existing methods. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 40, 2015. 1

Vikas Karade and Bhallamudi Ravi. 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. 1

Yoni Kasten, Daniel Doktofsky, and Ilya Kovler. End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pp. 123–133. Springer, 2020. 1, 2, 3, 4, 5, 6, 7, 9

Tanja Lerchl, Malek El Husseini, Amirhossein Bayat, Anjany Sekuboyina, Luis Hermann, Kati Nispel, Thomas Baum, Maximilian T Löffler, Veit Senner, and Jan S Kirschke. Validation of a patient-specific musculoskeletal model for lumbar load estimation generated by an automated pipeline from whole body ct. *Frontiers in Bioengineering and Biotechnology*, 10, 2022. 6

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 5

Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship: Some ml papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17 (1):45–77, 2019. 3

Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, 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. 4

Chan Li Jin Melissa, Frank Guan Yunqing, Luis Lanca, Jenni Ahjonpalo, Jenna Perälahti, Teemu Iivarinen, Vesa Ollikainen, and Eija Metsälä. Using convolutional neural network in 3d reconstruction from 2d radiographs-a scoping review. *Kliininen Radiografiatiede*, pp. 13, 2021. 1

D Mitton, C Landry, S Veron, Wata Skalli, F Lavaste, and Jacques A De Guise. 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. 1

Megumi Nakao, Fei Tong, Mitsuhiro Nakamura, and Tetsuya Matsuda. Image-to-graph convolutional network for deformable shape reconstruction from a single projection image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 259–268. Springer, 2021. 1, 2

Mark J Pearcy. Stereo radiography of lumbar spine motion. *Acta Orthopaedica Scandinavica*, 56 (sup212):1–45, 1985. 1

Charles G Peterfy, Erika Schneider, and M Nevitt. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and cartilage*, 16(12):1433–1441, 2008. 4

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021. 3

Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical image analysis*, 73:102166, 2021. 4

Ryoya Shiode, Mototaka Kabashima, Yuta Hiasa, Kunihiro Oka, Tsuyoshi Murase, Yoshinobu Sato, and Yoshito Otake. 2d–3d reconstruction of distal forearm bone from actual x-ray images of the wrist using convolutional neural networks. *Scientific Reports*, 11(1):1–12, 2021. 1, 2, 3, 4, 5, 7, 9

Marco A Siccardi and Bruno Bordoni. Anatomy, abdomen and pelvis, perineal body. 2019. 7

Weinan Song, Yuan Liang, Jiawei Yang, Kun Wang, and Lei He. Oral-3d: reconstructing the 3d structure of oral cavity from panoramic x-ray. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 566–573, 2021. 1

Chung-Ha Suh. The fundamentals of computer aided x-ray analysis of the spine. *Journal of biomechanics*, 7(2):161–169, 1974. 1

Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3405–3414, 2019. 5

Jeroen Van Houtte, Emmanuel Audenaert, Guoyan Zheng, and Jan Sijbers. Deep learning-based 2d/3d registration of an atlas to biplanar x-ray images. *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–10, 2022. 1

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):1–8, 2022. 3

J. Wu. Itk-based implementation of two-projection 2d/3d registration method with an application in patient setup for external beam radiotherapy. *The Insight Journal*, 12 2010. doi: 10.54294/6f280b. 3

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 5