Self-Supervised Multiview Xray Matching

Mohamad Dabboussi[1,2], Malo Huard[2], Yann Gousseau[1], and Pietro Gori[1]

[1] Telecom Paris, Palaiseau, France
[2] Milvue, Paris, France

Abstract. Accurate interpretation of multi-view radiographs is crucial for diagnosing fractures, muscular injuries, and other anomalies. While significant advances have been made in AI-based analysis of single images, current methods often struggle to establish robust correspondences between different X-ray views, an essential capability for precise clinical evaluations. In this work, we present a novel self-supervised pipeline that eliminates the need for manual annotation by automatically generating a many-to-many correspondence matrix between synthetic X-ray views. This is achieved using digitally reconstructed radiographs (DRR), which are automatically derived from unannotated CT volumes. Our approach incorporates a transformer-based training phase to accurately predict correspondences across two or more X-ray views. Furthermore, we demonstrate that learning correspondences among synthetic X-ray views can be leveraged as a pretraining strategy to enhance automatic multi-view fracture detection on real data. Extensive evaluations on both synthetic and real X-ray datasets show that incorporating correspondences improves performance in multi-view fracture classification.

Keywords: Multi-view X-ray · DRR · Many-to-Many Correspondence · Fracture detection.

1 Introduction

Accurate diagnosis in radiology often relies on the complementary information provided by multiple X-ray views that are acquired from various angles to ensure a comprehensive evaluation. In practice, radiologists examine all these views to confirm the presence and extent of lesions, thus increasing diagnostic confidence.
Motivation Multi-view imaging is essential in radiology [1,3,20], as each X-ray projection provides unique information that aids in detecting subtle abnormalities and improving overall evaluation. However, the process is time-consuming, requires expertise, and is prone to errors. These challenges highlight the need for automated methods capable of effectively handling multi-view data to support clinical decision-making.
Problem Statement While deep learning achieved impressive results in single-view medical imaging, extending it to multi-view X-ray interpretation remains challenging. Progress is hindered by two key issues: (1) the scarcity of annotated multi-view datasets, especially beyond chest X-rays, and (2) the complexity of
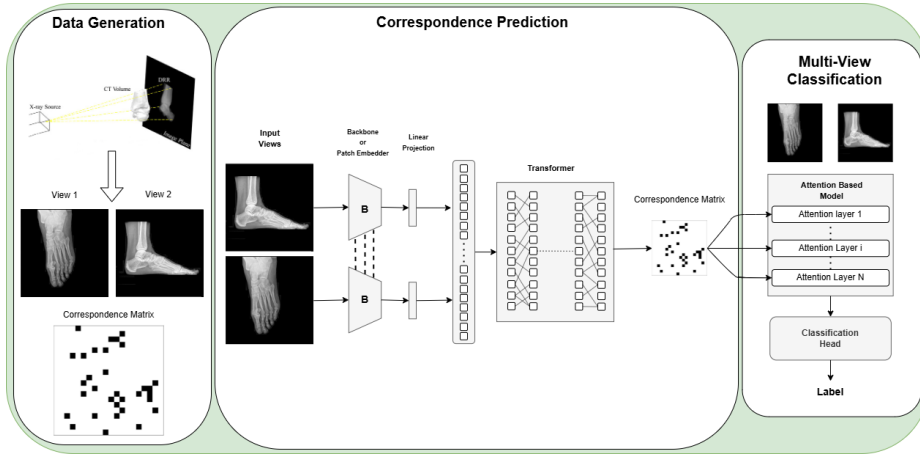
Fig. 1. End-to-end self-supervised pipeline for predicting X-ray correspondences and multi-view fracture classification. (Left) From a CT volume we generate multiple 2D DRRs (digitally reconstructed radiographs) with a correspondence matrix linking matching points. (Middle) A backbone network extracts features from each view, processed by a Transformer to predict a Correspondence Matrix via attention mechanisms. (Right) The correspondence guides attention for fracture detection.

establishing correspondences between views. Traditional image matching techniques focus on one-to-one mappings in natural images, whereas X-rays involve complex, many-to-many relationships due to the cumulative nature of pixel intensities along the X-ray path (see Fig. 2). This makes it challenging to determine whether abnormalities across views correspond to the same pathology. There is a need for an automated method to estimate accurate multi-view correspondences without extensive manual annotation, ultimately aiding radiologists and improving lesion/fracture detection.

To this end, our work (see Fig. 1 for a visual pipeline) introduces the following **contributions**:

- **DRR-based Correspondence Generation:** We generate paired simulated X-rays views and patch-level correspondence matrices from unannotated CT volumes, thereby eliminating the need for large annotated X-ray datasets.

- **Self-supervised Pre-training:** We leverage the correspondence prediction task as a self-supervised method to enhance multi-view feature learning.

- **Transformer Integration:** We incorporate correspondence information within transformer-based architectures to improve fracture classification performance.

2   Related Works

DRR-based Dataset Generation Digitally Reconstructed Radiographs (DRR) synthesize X-ray images from CT volumes. Early ray-tracing methods [10, 19]
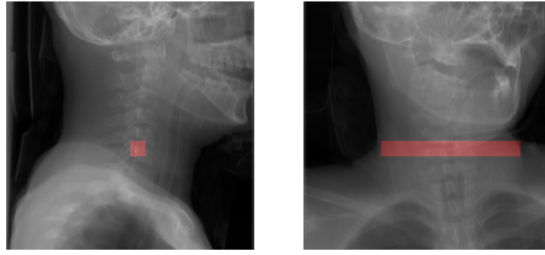
Fig. 2. Many-to-many matching between two X-ray views of the same subject. The red patch in the left view corresponds to the red bar in the right view, illustrating multiple matching pixels across views.

optimized efficiency, while recent deep learning approaches [5,23] improved realism. DRRs have been used for: (1) 3D CT reconstruction from limited X-ray views [4], (2) 2D/3D image registration [18,25,26], and (3) generating labeled datasets for segmentation using manually labeled 3D CT volumes [11,12,27]. Unlike prior work, we leverage DRR to automatically generate correspondence annotations from *unannotated* CT volumes, enabling large-scale, annotated (i.e., correspondences), multi-view X-ray datasets across diverse anatomical regions.

Image Matching and Correspondence Estimation   Traditional feature-based image matching [9,14] has been largely replaced by deep learning approaches. Sparse keypoint methods [13,17] and dense matching techniques [2,7,22] perform well in natural image tasks but assume one-to-one pairwise matching, which suffices for tasks involving two views with homographies, such as camera pose estimation. Multi-view X-ray imaging, however, requires many-to-many mappings due to anatomical transparency and overlapping structures. Existing methods struggle in this context, highlighting the need for specialized solutions in X-ray correspondence estimation.

To address these challenges, we propose a self-supervised pipeline that generates from unannotated CT volumes a large, diverse dataset of synthetic X-rays with automatically derived correspondence matrices, eliminating the need for manual annotations. Unlike traditional pairwise matching methods, our approach learns many-to-many correspondences across multiple views, enabling a more comprehensive understanding of their complex spatial relationships and advancing automated multi-view X-ray analysis.

3   Method

We introduce a novel approach for generating correspondence ground truth using Digitally Reconstructed Radiographs (DRR) from unannotated CT scans. Synthetic X-ray views are first created along with their corresponding correspondence matrices. These are then used as supervision for training a deep-learning model, which takes two synthetic X-ray views as input and predicts their correspondence matrix.

To evaluate the effectiveness of our approach, we conduct experiments on both synthetic and real multi-view X-ray datasets, demonstrating the effectiveness of correspondence learning not only as a pretraining strategy but also as a mechanism for attention guidance in multi-view fracture classification.

3.1   Correspondence Ground Truth Generation

Let $V \in \mathbb{R}^{H \times W \times L}$ denote a CT volume consisting of N voxels. For each non-air voxel v at position (x, y, z), obtained by thresholding the CT volume, we project it onto two distinct view directions using the Joseph method [10] to generate DRRs. This results in two projection matrices, $P_1^v, P_2^v \in \mathbb{R}^{H \times W}$, which encode the accumulated intensity values along rays traced from the X-ray source to the detector plane (see Fig. 4 for a visual explanation). We then define their flattened representations as: $\mathbf{p}_1^v = flat(P_1^v) \in \mathbb{R}^{HW}$ and $\mathbf{p}_2^v = flat(P_2^v) \in \mathbb{R}^{HW}$, where $flat(\cdot)$ denotes the flattening of a matrix into a vector. The voxel-specific correspondence matrix is then given by the outer product: $\mathbf{C}_{1,2}^v = \mathbf{p}_1^v \mathbf{p}_2^{v\top} \in \mathbb{R}^{HW \times HW}$. The final correspondence matrix is obtained by taking the element-wise maximum over all voxels: $\mathbf{C}_{1,2} = \max_v \left(\mathbf{C}_{1,2}^v \right)$.

To address computational challenges posed by high-resolution CT volumes (e.g., $256 \times 256 \times 256$), we perform patch-level (and not pixel-level) correspondence estimation (see Fig. 3). Specifically, we downsample the volume by a factor of $k = 16$, thereby computing the correspondence matrix at a coarser, patch-wise resolution.
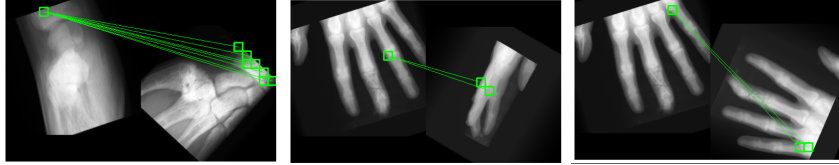

Fig. 3. Visualization of generated hand correspondences annotations.

3.2   Correspondence Prediction

We formulate correspondence prediction as a similarity assessment between features extracted from patches across different views. A pretrained backbone network, denoted as f, extracts feature maps from each image, which are then projected into a lower-dimensional embedding space. The spatial grid of these feature maps is flattened into a sequence of patch embeddings, where each embedding represents a distinct image region. Afterwards, patch embeddings from multiple views are concatenated along the sequence dimension and processed by a transformer module. This module employs self-attention to capture both
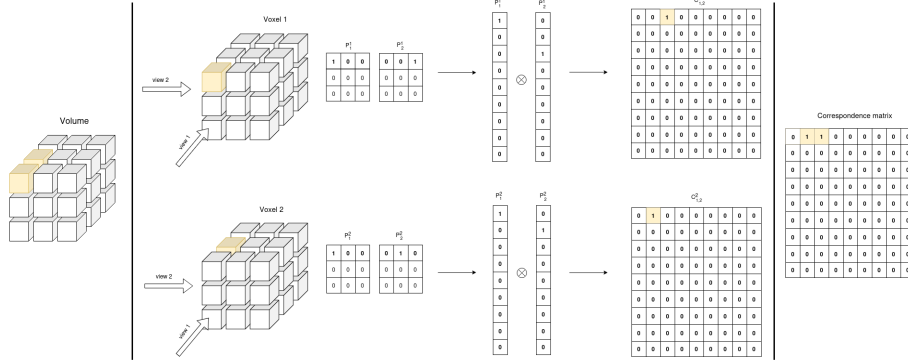
Fig. 4. Generating a correspondence matrix for a $3 \times 3 \times 3$ volume with a highlighted $1 \times 1 \times 2$ cube (yellow) using two orthogonal views. Each row shows the projection of a yellow voxel v onto both views. The resulting projection matrices, P_1^v and P_2^v, are flattened, and their outer product forms $C_{1,2}^v$. The final correspondence matrix is obtained by taking the maximum over all $C_{1,2}^v$ matrices.

intra- and inter-view interactions, effectively encoding correspondence information. Correspondences are then determined by computing a normalized dot product between feature pairs, forming a correspondence matrix. This approach naturally extends to multi-view scenarios with more than two images.

3.3   Pre-training through correspondence prediction

As we will show in Section 4, we can use correspondence prediction as a self-supervised pretraining step for an auxiliary multi-view X-ray classification task. This approach encourages the model to learn robust features that capture shared anatomical information while leveraging the redundancy in multi-view data.

3.4   Attention Guidance in Multi-View X-ray classification

To leverage cross-view correspondence information in our multi-view X-ray classification framework, we can also integrate the correspondence matrix directly into the transformer attention mechanism.

Specifically, we concatenate the N patch tokens from both views A and B into a 2N×D sequence, producing a 2N×2N attention map split into four N×N blocks: AA, BB (self-attention), AB, and BA (cross-attention). We add the learned correspondence matrix to AB and its transpose to BA as a soft bias before the softmax in each attention layer, guiding cross-view focus.

Let Q and K denote the query and key matrices, respectively, and let d be the dimensionality of each attention head. The standard scaled dot-product attention is computed as $A = \frac{QK^\top}{\sqrt{d}}$ (Eq. 1). Given a correspondence matrix C, the attention scores are adjusted as $A' = A + \alpha\, C$ (Eq. 2), where α is a learnable

scaling parameter. Finally, the modified scores A' are normalized via the softmax function to yield the attention probabilities.

4   Experiments

4.1   Datasets and Implementation Details

Train Simulated Dataset We generated correspondence matrices using our proposed method. In particular, 207,600 data samples were generated from 175 CT volumes by varying projection angles, the distance between the volume and the source, as well as by extracting appropriate crops from the CT volumes.

Test Simulated X-ray Correspondence Dataset From 42 CT volumes of extremity, not seen during training, 16,920 pairs of views along with their corresponding correspondence matrices were generated using the same variation parameters as during training.

Real X-ray Correspondence Dataset This dataset comprises 347 samples, each with two views. We applied a 60%/10%/30% split for training, validation, and testing. Each sample includes only three annotated positive correspondences, along with 100 negative correspondences.

MURA Public Dataset The MURA dataset [16] comprises studies from various anatomical regions. For our purposes, we selected studies of the elbow, forearm, hand, and wrist, which correspond to the regions on which the correspondence model was trained. The final dataset includes 5,511 studies. Each study has multiple views, we used 2 views in our experiments.

Private Dataset This dataset mainly contains multiview X-rays of the hand, forearm, foot, and knee. It has a total of 5,653 studies. Each study has multiple views, we used 2 views in our experiments.

Implementation Details Our framework processes two input views, each of size 256×256 pixels. We use a pre-trained ResNet-50 [8] backbone for feature extraction, followed by a transformer with Rotary Positional Encoding (RoPE [21]) to capture spatial relationships between patches. Training utilizes the Adam optimizer with a cosine annealing learning rate scheduler. The initial learning rate is set to 1×10^{-4} for pre-training on simulated correspondence data and reduced to 1×10^{-5} when fine-tuning on partially annotated real data. The model is trained with a batch size of 16. For loss, we use mean squared error (MSE) for the correspondence task and binary cross-entropy for classification. Our classification model uses a ViT-S (6 layers, 4 heads, hidden size 384). All model modules were pre-trained on synthetic DRR correspondence data and subsequently fine-tuned for the binary fracture classification task. Each transformer layer uses a distinct learnable α for attention guidance. Importantly, our attention-bias module is modular and can be integrated into other attention-based architectures, making it broadly applicable to classification, segmentation, and detection tasks in multi-view settings. Data augmentation techniques, including random adjustments to brightness, contrast, and color inversion, are applied to enhance model robustness.

4.2   Results and Discussion

Table 1. Model Performance on Correspondence Simulated Test Dataset.

Attention Model	Message Pass	MSE	Precision	Recall	AP
–	–	2.25×10^{-3}	49.1	68.0	54.7
Superglue Module	✓	1.33×10^{-3}	70.3	75.6	80.1
LoFTR Module	✓	1.24×10^{-3}	74.2	78.1	83.5
Standard Transformer	✓	1.19×10^{-3}	75.5	79.4	84.1
Standard Transformer	–	$\mathbf{1.09 \times 10^{-3}}$	**77.0**	**81.5**	**85.0**

Correspondence prediction - Simulated data Table 1 summarizes the performance of various models for our patch-level correspondence prediction task on a simulated test dataset. We use ResNet-50 as the backbone for feature extraction and compute the correspondence matrix using normalized dot product. Our experiments compare different attention modules: graph-based message passing methods (e.g., SuperGlue [17] and LoFTR [22]) and a standard transformer attention module. The evaluation metrics include mean squared error (MSE), precision, recall, and average precision (AP). Notably, our approach, the Standard Transformer [24] without message passing, achieved the best performance across all metrics, with the lowest MSE and the highest AP. This suggests that the transformer's flexible attention mechanism enables more effective inter-patch communication compared to the more constrained graph-based approaches. Since our goal is to predict correspondences at the patch level, acknowledging that abnormalities typically span groups of pixels rather than isolated pixels, it is natural to focus on patches. In contrast, a CNN-only baseline without any attention mechanism performed considerably worse, highlighting the importance of attention mechanisms in learning robust patch-level correspondences.

Table 2. Model Performance on the Correspondence Test Dataset of Real X-ray.

Method	Backbone	Model	Precision	Recall	AP
Zero Shot	Res50	LoFTR module	32.5	2.7	9.4
	Res50	LoFTR* module	28.3	26.3	15.9
	DinoV2-G	–	3.0	15.2	1.7
Fine-tuning	Res50	LoFTR* module	36.8	18.0	16.1
	DinoV2-G	Multi Layer Perceptron	55.23	51.6	42.2
Pre-train + Fine-tuning	Res50	Standard Transformer	**72.3**	**87.1**	**83.8**

Correspondence prediction - Real data Table 2 evaluates our method for correspondence matching on a real X-ray dataset with partial annotations, comparing it against existing approaches in both zero-shot and fine-tuned settings. The goal is to establish reliable point correspondences with minimal annotation. In the zero-shot setting, the pre-trained LoFTR model [22] with a ResNet50 backbone struggles due to domain shift and its one-to-one matching strategy, resulting in low recall (2.7) and modest AP (9.4). We introduce a variant LoFTR*, which incorporates a normalized dot-product and thresholding correspondence head for multi-to-multi matching, slightly improving recall (26.3) and AP (15.9), though the domain gap remains significant. Fine-tuning LoFTR* improves performance (AP: 16.1), but a stronger baseline is obtained by freezing a 1.1B-parameter DinoV2-G model [15] and fine-tuning an MLP head, achieving much better results (AP: 42.2). Our approach, a transformer-based model pre-trained and fine-tuned with only 24M parameters, outperforms all other methods, achieving the highest AP (83.8).

Table 3. Classification Performance on the MURA and the Private Dataset.

	Correspondence Pretraining	Attention Guidance	Fusion	Accuracy	Precision	Recall	Kappa
MURA	–	–	Single	73.3 ± 0.3	75.4 ± 0.2	65.4 ± 0.2	0.45
	–	–	Late	78.7 ± 0.2	83.3 ± 0.1	67.4 ± 0.2	0.56
	–	–	Early	75.8 ± 0.2	78.5 ± 0.1	65.4 ± 0.2	0.49
	✓	–	Early	$\underline{80.1 \pm 0.1}$	$\underline{83.4 \pm 0.1}$	$\underline{73.6 \pm 0.2}$	$\underline{0.58}$
	✓	✓	Early	$\mathbf{80.6 \pm 0.1}$	$\mathbf{84.8 \pm 0.1}$	$\mathbf{74.4 \pm 0.2}$	**0.59**
Private	–	–	Single	68.8 ± 0.2	53.6 ± 0.2	41.7 ± 0.2	0.30
	–	–	Late	74.2 ± 0.2	59.5 ± 0.2	$\underline{50.1 \pm 0.2}$	0.36
	–	–	Early	71.1 ± 0.1	54.4 ± 0.1	44.2 ± 0.1	0.32
	✓	–	Early	$\underline{75.0 \pm 0.1}$	$\underline{59.7 \pm 0.1}$	49.2 ± 0.2	$\underline{0.37}$
	✓	✓	Early	$\mathbf{76.2 \pm 0.1}$	$\mathbf{59.9 \pm 0.1}$	$\mathbf{52.6 \pm 0.2}$	**0.39**

Multi-View X-ray classification Table 3 evaluates the impact of correspondence pretraining and attention guidance on multi-view X-ray classification. This experiment demonstrates the benefits of pretraining a model on a correspondence task before applying it to classification, as well as the advantages of using the correspondence mask to guide attention. We used a ViT model [6] trained on both public and private datasets. We compare early fusion, where patch embeddings from multiple views are concatenated before input to the transformer, and late fusion, where scores are aggregated after independent processing. Multi-view fusion significantly outperforms single-view methods, and with correspondence pretraining, early fusion surpasses late fusion.

Pretraining on correspondence further boosts performance, despite all models being initialized with ImageNet weights. Notably, early fusion with pretraining improves accuracy from 75.8% to 80.1% on the public dataset. Additionally,

incorporating the correspondence matrix as an attention bias further enhances results, yielding the highest accuracy (80.6% public, 76.2% private).

5 Conclusion

We introduced a new framework for multi-view X-ray analysis that leverages self-supervised correspondence learning to improve both correspondence matching on real X-ray images and multi-view fracture detection. Our experiments show that pretraining on correspondence significantly enhances classification accuracy, especially when integrating correspondence information into transformer-based architectures. As demonstrated in our fracture classification task, the no-annotation correspondence method we proposed opens up numerous use cases in multi-view X-ray tasks. This work provides a novel approach that can be applied to a variety of X-ray tasks, advancing the field of multi-view medical image analysis.

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

References

1. Brandser, E.A., Berbaum, K.S., Dorfman, D.D., Braksiek, R.J., El-Khoury, G.Y., Saltzman, C.L., Marsh, J.L., Clark, W.A.: Contribution of individual projections alone and in combination for radiographic detection of ankle fractures. AJR Am. J. Roentgenol. **174**(6), 1691–1697 (Jun 2000)
2. Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., Mckinnon, D., Tsin, Y., Quan, L.: Aspanformer: Detector-free image matching with adaptive span transformer. In: European Conference on Computer Vision. pp. 20–36. Springer (2022)
3. Cheung, G.C., Lever, C.J., Morris, A.D.: X-ray diagnosis of acute scaphoid fractures. J. Hand Surg. Br. **31**(1), 104–109 (Feb 2006)
4. Corona-Figueroa, A., Frawley, J., Bond-Taylor, S., Bethapudi, S., Shum, H.P., Willcocks, C.G.: Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In: 2022 44th annual international conference of the IEEE engineering in medicine & Biology society (EMBC). pp. 3843–3848. IEEE (2022)
5. Dhont, J., Verellen, D., Mollaert, I., Vanreusel, V., Vandemeulebroucke, J.: Realdrr–rendering of realistic digitally reconstructed radiographs using locally trained image-to-image translation. Radiotherapy and Oncology **153**, 213–219 (2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024)

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

9. Herbert, B.: Surf: Speeded up robust features. Computer vision and image understanding **110**(3), 346–359 (2008)

10. Joseph, P.M.: An improved algorithm for reprojecting rays through pixel images. IEEE transactions on medical imaging **1**(3), 192–196 (1982)

11. Kasten, Y., Doktofsky, D., Kovler, I.: End-to-end convolutional neural network for 3d reconstruction of knee bones from bi-planar x-ray images. In: Machine Learning for Medical Image Reconstruction: Third International Workshop, MLMIR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3. pp. 123–133. Springer (2020)

12. Killeen, B.D., Wang, L.J., Zhang, H., Armand, M., Taylor, R.H., Osgood, G., Unberath, M.: Fluorosam: A language-aligned foundation model for x-ray image segmentation. arXiv preprint arXiv:2403.08059 (2024)

13. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17627–17638 (October 2023)

14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)

15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

16. Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L., et al.: Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957 (2017)

17. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)

18. Shrestha, P., Xie, C., Shishido, H., Yoshii, Y., Kitahara, I.: X-ray to ct rigid registration using scene coordinate regression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 781–790. Springer (2023)

19. Siddon, R.L.: Prism representation: a 3d ray-tracing algorithm for radiotherapy applications. Physics in Medicine & Biology **30**(8), 817 (1985)

20. Southall, C., Friebel, T.R., Gardiner, S., Nikkhah, D.: Radiographic views for hand fractures - call for three-view national UK guidelines - a quality improvement study. Int. J. Surg. **109**(7), 1919–1922 (Jul 2023)

21. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

22. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8922–8931 (June 2021)

23. Unberath, M., Zaech, J.N., Lee, S.C., Bier, B., Fotouhi, J., Armand, M., Navab, N.: Deepdrr–a catalyst for machine learning in fluoroscopy-guided procedures. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. pp. 98–106. Springer (2018)

24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

25. Wu, H., Zhang, J., Fang, Y., Liu, Z., Wang, N., Cui, Z., Shen, D.: Multi-view vertebra localization and identification from ct images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 136–145. Springer (2023)
26. Zhang, B., Faghihroohi, S., Azampour, M.F., Liu, S., Ghotbi, R., Schunkert, H., Navab, N.: A patient-specific self-supervised model for automatic x-ray/ct registration. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 515–524. Springer Nature Switzerland, Cham (2023)
27. Zhang, Y., Miao, S., Mansi, T., Liao, R.: Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 599–607. Springer (2018)