Cross-view Deformable Transformer for Non-displaced Hip Fracture Classification from Frontal-Lateral X-ray Pai

Zhonghang Zhu, Qichang Chen, Lequan Yu, Lianxin Wang, Defu Zhang, Baptiste Magnier, Liansheng Wang

Cross-view Deformable Transformer for Non-displaced Hip Fracture Classification from Frontal-Lateral X-ray Pair

Zhonghang Zhu[1], Qichang Chen[1], Lequan Yu[2], Lianxin Wang[3], Defu Zhang[1], Baptiste Magnier[4], and Liansheng Wang[1](✉)

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China, {zzhonghang,qcchen}@stu.xmu.edu.cn, {lswang,dfzhang}@xmu.edu.cn

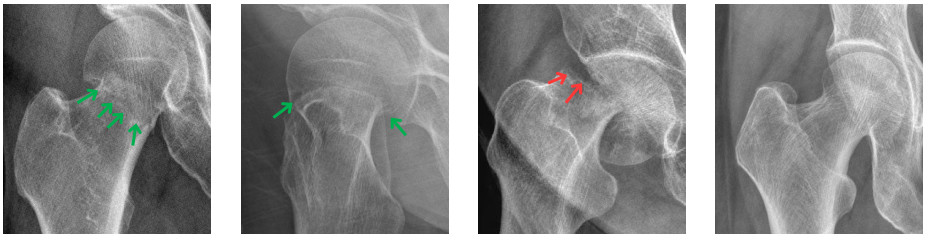
[2] Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong SAR, China, lqyu@hku.hk

[3] Department of Orthopedics, The First Affiliated Hospital of Xiamen University, Xiamen, China, dr_shepherd@sina.com

[4] Euromov Digital Health in Motion, Univ Montpellier, IMT Mines Ales, Ales, France, baptiste.magnier@mines-ales.fr

Abstract. Hip fractures are a common cause of morbidity and mortality and are usually diagnosed from the X-ray images in clinical routine. Deep learning has achieved promising progress for automatic hip fracture detection. However, for fractures where displacement appears not obvious (*i.e.*, non-displaced fracture), the single-view X-ray image can only provide limited diagnostic information and integrating features from cross-view X-ray images (*i.e.*, Frontal/Lateral-view) is needed for an accurate diagnosis. Nevertheless, it remains a technically challenging task to find reliable and discriminative cross-view representations for automatic diagnosis. First, it is difficult to locate discriminative task-related features in each X-ray view due to the weak supervision of image-level classification labels. Second, it is hard to extract reliable complementary information between different X-ray views as there is a displacement between them. To address the above challenges, this paper presents a novel cross-view deformable transformer framework to model relations of critical representations between different views for non-displaced hip fracture identification. Specifically, we adopt a deformable self-attention module to localize discriminative task-related features for each X-ray view only with the image-level label. Moreover, the located discriminative features are further adopted to explore correlated representations across views by taking advantage of the query of the dominated view as guidance. Furthermore, we build a dataset including 768 hip cases, in which each case has paired hip X-ray images (Frontal/Lateral-view), to evaluate our framework for the non-displaced fracture and normal hip classification task.

Keywords: Hip fracture diagnosis · X-ray image · Deformable transformer · Cross-view correspondence.



(a) Non-displaced Frontal (b) Non-displaced Lateral (c) Displaced Frontal (d) Normal Frontal

Fig. 1. Comparisons of non-displaced/displaced hip fracture and normal hip X-ray images. The fracture regions are marked by green arrows and red arrows for non-displaced/displaced fractures, respectively.

1 Introduction

Hip fractures represent a life-changing event and carry a substantial risk of decreased functional status and death, especially in elderly patients [17]. Usually, they are diagnosed from X-ray images in clinical practice. Currently, proper X-ray fracture identification relies on the manual observation of board-certified radiologists, which leads to increased workload pressures to radiologists. However, accurate and timely diagnosis of hip fractures is critical, especially in emergency situations such as non-displaced hip fractures [13]. Therefore, automated X-ray image classification is of great significance to support the clinical assistant diagnosis.

Recently, Deep Learning (DL) methods for radiography analysis have gained popularity and shown promising results [4,18,7,14], which aims to distinguish normal radiography or prioritize urgent/critical cases with the goal of reducing the radiologist workload or improving the reporting time. For example, a triaging pipeline based on the urgency of exams has been proposed in [1] and Tang *et al.* [21] compared different DL models applied to several public chest radiography datasets for distinguishing abnormal cases. However, these works only focus on single-view radiography analysis. When the fracture displacement in the X-ray image is not apparent,*i.e.*, a non-displaced hip fracture as shown in Fig. 1, these methods may fail in extracting enough fracture features represented by a ridge [20] in the image and result in misdiagnosis. Therefore, it is necessary to develop cross-view learning approaches to diagnose fracture from paired views (Frontal/Lateral-images), which have been demonstrated to provide complementary features to promote the diagnostic performance [3,10]. Recent studies have been investigated for cross-view learning of X-ray images, which aims to exploit the value of paired X-ray images and fuse them to get a comprehensive anatomical representation for diagnosis [19,22,15,2,5]. However, these methods do not consider cross-view feature relations which is a quite important issue for accurate cross-view feature fusion.

Since the introduction of vision transformer models [9], more researches have been developed in the tokenization process and relation modeling among tokens

in an image [8,16]. Recently, deformable self-attention has been proposed to refine visual tokens [6,25,23], which is powerful in focusing on relevant regions and capturing more informative features. Motivated by this, we propose a novel cross-view deformable transformer framework for hip fracture detection from cross-view X-ray images. Firstly, deformable self-attention modules are utilized to localize reliable task-related features of each view. Secondly, the dominated-view characteristics are used to explore informative representations in the other view for effective feature fusion of cross-view X-ray images. Specifically, our contributions are three folds:

1. We propose a cross-view deformable transformer framework for non-displaced hip fracture classification, in which we take advantage of discriminative features of Frontal-view as a guidance to localize informative representations of Lateral-view for cross-view feature fusion.
2. For each view, we adopt the deformable self-attention module to select pivotal tokens in a data-dependent way.
3. We build a new non-displaced hip fracture X-ray dataset which includes both Frontal and Lateral views for each case to valid the proposed method. Our approach surpasses the state of the art in accuracy by over 1.5%.

2 Method

2.1 Overview

The detailed architecture of the proposed cross-view deformable attention framework is shown in Fig. 2. To model the relations among features across different views, the framework is designed as a joint of two view-specific deformable transformer branches with four stages. For each view-specific branch, the input image is firstly processed by shifted window attention modules presented in the left of Fig. 2 to aggregate information locally, followed by the last two stages to model the global relations among the locally augmented tokens with deformable self-attention modules. In the last two stages, the query features of the Frontal-view are adopted as the guidance to detect the relations among the Lateral-view tokens. The detailed design of each component is introduced below.

2.2 View-specific Deformable Transformer Network

To discover the task-related regions of each view, the view-specific branch is designed as a deformable transformer network consisting of four stages. In each branch, the first two stages explore the local representations of the input images with shift-window attention modules, followed by the last two stages exploit local tokens relation using deformable self-attention modules. Specifically, our framework takes an image of size $H \times W \times 3$ as input. After the first two stages, the input image will be embedded into feature maps $f_{layer3} \in H/4 \times W/4 \times C$, where the C denotes the channel number. The f_{layer3} will be passed to a query

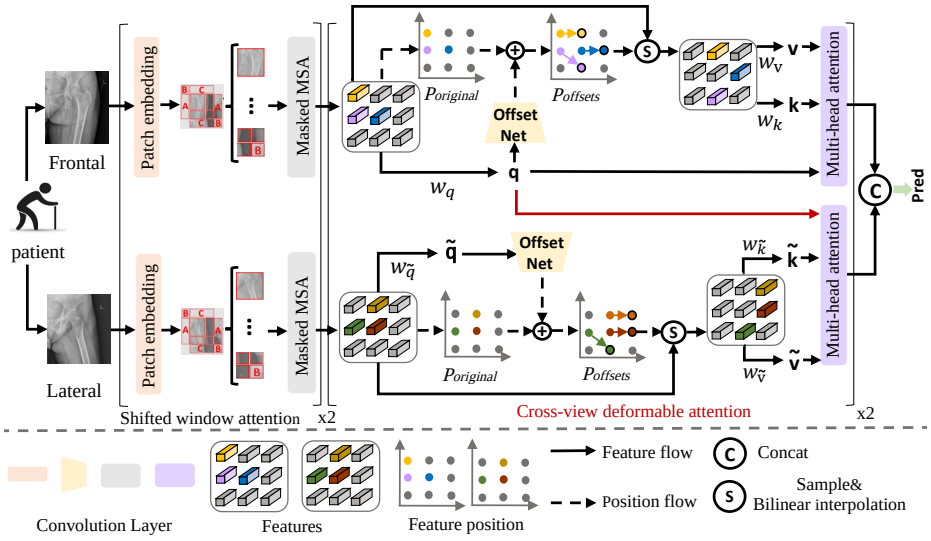


Fig. 2. Illustration of the proposed framework depicted in two view-specific branches with four stages. A pair of X-ray images, *i.e.*, Frontal and Lateral images are fed into two branches, respectively. The input images are processed with shifted window attention modules to aggregate discriminative local features (first two stages), while deformable self-attention modules are utilized to model the relations among tokens (last two stages). Moreover, Frontal queries are passed to model relations among Lateral features for cross-view deformable attention. The $p_{original}$ represents the original feature position of each view, while $p_{offsets}$ denotes the position offsets. The $W_{q/\tilde{q}}$, $W_{k/\tilde{k}}$ and $W_{v/\tilde{v}}$ are projection matrices for queries, keys and values, respectively.

projection network W_q, which is a light network to obtain the query feature maps $f_{layer3;q}$. Moreover, a uniform grid $p_{original} \in \mathbb{R}^{H/4\times W/4\times 2}$ is generated as a position reference of points in f_{layer3}. The values of $p_{original}$ are linearly spaced and normalized to 2D coordinates range in $(-1,-1),\cdots,(1,1)$, in which the $(-1,\cdot),\cdots,(1,\cdot)$ and the $(\cdot,-1),\cdots,(\cdot,1)$ refers to the horizontal and vertical coordinates for reference points respectively. In the meanwhile, reference points offset $p_{offsets} \in \mathbb{R}^{H/4\times W/4\times 2}$ are generated from the $f_{layer3;q}$ by a light offset network consisted of two convolutional layers followed normalization layer, which are also normalized into $(-4/H,-4/W),\cdots,(4/H,4/W)$. The shifted position of points in f_{layer3} are calculated as $p = \psi(p_{original} + a(p_{offsets}))$, where $a(\cdot)$ is a function (*i.e.*, $4\tanh(\cdot)$) to prevent the offset from becoming too large and $\psi(p_x, p_y) = (H * p_x, W * p_y)$. Then the deformed features of each point are sampled at the shifted position, which could be denoted as $\bar{f} = S(f, p)$, where S represents a bilinear interpolation function. Therefore, the deformed multi-head self-attention module with M heads can be described as:

$$q = fW_q,\ \bar{k} = \bar{f}W_k,\ \bar{v} = \bar{f}W_v, \tag{1}$$

$$z^m = \sigma(q^{(m)}\bar{k}^{(m)T}/\sqrt{d})\bar{v}^{(m)}, \quad m = 1,\cdots,M, \tag{2}$$

$$z = Concat\left(z^1, \cdots, z^M\right) W_o, \tag{3}$$

where $\sigma(\cdot)$ denotes the softmax function, and d is the dimension of each head. $z^{(m)}$ is the embedding output from the m-th attention head, and $\{q^{(m)}, \bar{k}^{(m)}, \bar{v}^{(m)}\} \in \mathbb{R}^{N\times d}$ represents query, deformed key and value embeddings, respectively. Also, W_q, W_k, W_v, W_o are the projection networks. Features passed to the 4th stage are conducted a same operation as in 3rd stage with different feature dimensions.

2.3 Cross-view Deformable Transformer Framework

The proposed cross-view framework is consisted of two joint view-specific branches, with a pair of X-ray images (Frontal-view and Lateral-view) from the same patient taken as the input of two individual view-specific branches, respectively. These input images will be embedded into primary representations in the first stage of view-specific network, then these primary features will be sent to the second stage to get representations with larger receptive field. To observe correlations between Frontal-view and Lateral-view, we opt for a simple solution to share queries from the Frontal-view to model token relations of Lateral-view in a self-attention manner as the Frontal-view contains dominated diagnosis features [24]. In this way, the focused regions of the Lateral-view are determined by the discriminative features of the Frontal-view. So for the Lateral-view branch, the multi-head self-attention can be denoted as:

$$q_{fr} = f_{fr}W_{q;fr},\ \bar{k_{la}} = \bar{f_{la}}W_{k;la},\ \bar{v_{la}} = \bar{f_{la}}W_{v;la}, \tag{4}$$

$$z_{la}^m = \sigma(q_{fr}^{(m)}\bar{k_{la}}^{(m)T}/\sqrt{d})\bar{v_{la}}^{(m)}, \quad m = 1, \cdots, M, \tag{5}$$

$$z_{la} = Concat\left(z_{la}^1, \cdots, z_{la}^M\right) W_{o;la}, \tag{6}$$

in which $(\cdot)_{fr}$ and $(\cdot)_{la}$ represent the features of Frontal-view and Lateral-view, respectively. While for the Frontal-view branch, the multi-head self-attention can be denoted as:

$$q_{fr} = f_{fr}W_{q;fr},\ \bar{k_{fr}} = \bar{f_{fr}}W_{k;fr},\ \bar{v_{fr}} = \bar{f_{fr}}W_{v;fr}, \tag{7}$$

$$z_{fr}^m = \sigma(q_{fr}^{(m)}\bar{k_{fr}}^{(m)T}/\sqrt{d})\bar{v_{fr}}^{(m)}, \quad m = 1, \cdots, M, \tag{8}$$

$$z_{fr} = Concat\left(z_{fr}^1, \cdots, z_{fr}^M\right) W_{o;fr}, \tag{9}$$

It is worth noting that view-specific reference points offset are derived from corresponding view-specific query feature maps which contain global view-specific position relations. By taking query feature maps from the Frontal-view as an informative clue, it makes sense to search relevant task-related features in the Lateral-view deformed values and keys embedding which are also discriminative features of Lateral-view. In this way, the cross-view transformer framework manages to localize task-related features in both views while exploring the cross-view related representations for feature aggregation. Then the final output can be denoted as $outputs = MLP(Concat(f_{fr}, f_{la}))$, where the f_{fr} and f_{la} represent the

output features of the last layer of the Frontal-view and Lateral-view branches, respectively. The *Concat* is a concatenation operation and *MLP* is a projection head consisted of two fully connected layers to generate logit predictions.

2.4 Technical Details

The view-specific model shares a similar pyramid structure with DTA-T [23]. The first stage consists of one shift-window block whose head number is set as 3, followed the second stage with one shift-window block whose head number is set as 6. We adopt three deformable attention block with 12 heads in the 3rd stage and one deformable attention block with 24 heads in the 4th stage. To optimize the whole framework, we calculate the cross entropy loss between the label and final output of the cross-view deformable transformer for training.

3 Experiment

3.1 Experiment Setup

Dataset. The dataset used in this study includes 768 paired hip X-ray images (329 non-displaced fractures, 439 normal hips) from 4 different manufacturers of radiologic data sources: GE Healthcare, Philips Medical Systems, Kodak and Canon. All the hip radiographs are collected and labeled by experts with non-displaced fractures or normal for classification task.

Implementation Details. For experiments of our dataset, we manually locate the hip region and crop a 224×224 image that is centered on the original hip region whose size is 400×600. The learning rate is set as 3e-3 for the end-to-end training of the framework with a batch size of 32. We adopt a 10-fold cross-validation and report the average performance of 10 folds. For each fold, we further divide the data (the other 9 folds) into a training set (90%) and a validation set (10%) and take the best model on the validation part for testing.

Evaluation Metric. We evaluate our method with Accuracy (Acc), *Precision*, *Recall* and F1 score. The *Precision* and *Recall* are calculated with one-class-versus-all-other-classes and then calculate F1 score $\left(F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}\right)$.

3.2 Experimental Results

Comparisons with the State of the Art. The proposed method is also compared with other cross-view fusion methods; results are reported in Table 1. *1) MVC-NET:* a network with back projection transposition branch to explicitly incorporate the spatial information from two views at the feature level. *2) DualNet:* an ensemble of two DenseNet-121 [12] networks followed a global average pooling operation of the final convolutional layer before a fully connected layer to simultaneously process multi-view images. *3) Auloss:* a DualNet regularized by auxiliary view-specific classification losses. *4) ResNet18-dual:* an

Table 1. Quantitative results (mean±standard deviation)% of different methods.

Method	Acc	*Precision*	*Recall*	F1 score
MVC-NET [26]	77.35±4.81	71.96±10.23	75.69±9.09	73.30±7.66
DualNet [19]	80.87±5.98	76.21±10.31	82.37±8.01	78.41±5.49
Auloss [10]	82.30±4.51	78.42±8.26	79.69±6.23	78.86±6.15
ResNet18-dual	88.41±2.86	95.87±3.27	76.14±8.29	84.58±4.61
Densenet-dual	90.36±3.42	94.37±3.11	82.20±7.95	87.66±4.60
Swin-dual	95.05±2.01	95.44±5.00	92.85±3.77	94.01±2.68
Lateral_swin	85.16±3.86	85.36±6.89	78.77±8.35	81.56±5.42
Lateral	84.37±4.43	84.51±7.78	78.01±9.47	80.68±6.41
Frontal_swin	91.67±3.53	94.12±5.22	85.55±8.64	89.33±5.34
Frontal	93.36±2.24	93.19±5.34	91.34±3.94	92.11±2.76
Ours w/o q	95.83±1.59	**96.10±4.82**	93.96±3.57	**96.36±1.31**
Ours	**96.48±1.83**	95.65±4.31	**96.22±3.69**	95.83±2.26

ensemble of two ResNet18 [11] networks, and the predicted results are generated by concatenating logit outputs from each ResNet18 network. *5) Densenet-dual:* an ensemble of two DenseNet networks, and the predicted results are generated by concatenating logit outputs from each DenseNet network. *6) Swin-dual:* an ensemble of two swin-transformer networks [16], and the predicted results are generated by concatenating logit outputs from each swin-transformer network.

As shown in Table 1, we compare our method to different dual frameworks. It can be observed that the proposed method achieves better performance than others (compare *Ours* with *ResNet18-dual*, *Densenet-dual* and *Swin-dual*), which demonstrates that the accuracy boost is due to the deformable transformer network and the feature interaction design not the increased backbone size. In addition, the *MVC-NET* shares a similar feature-level interaction motivation with *Ours*, and the 19.1% accuracy improvement indicates that our cross-view deformable attention gains better performance. Otherwise, we demonstrate the effectiveness of the deformable transformer network by comparing the *Our w/o q* to *Swin-dual*, as the only difference between these two frameworks is that the *Our w/o q* change the last two stages of *Swin-dual* to deformable transformer modules.

Ablation Study. We also conduct ablation experiments to validate the design of our proposed different components. We compare the following different settings. *1) Frontal:* take the Frontal image as input of the view-specific deformable transformer network to generate the prediction. *2) Frontal_swin:* take the Frontal image as input of the swin-transformer network to generate the prediction. *3) Lateral:* take the Lateral image as input of the view-specific deformable transformer network to generate the prediction. *4) Lateral_swin:* take the Lateral image as input of the swin-transformer network to generate the prediction. *5) Ours w/o q:* the proposed framework without cross-view deformable attention.

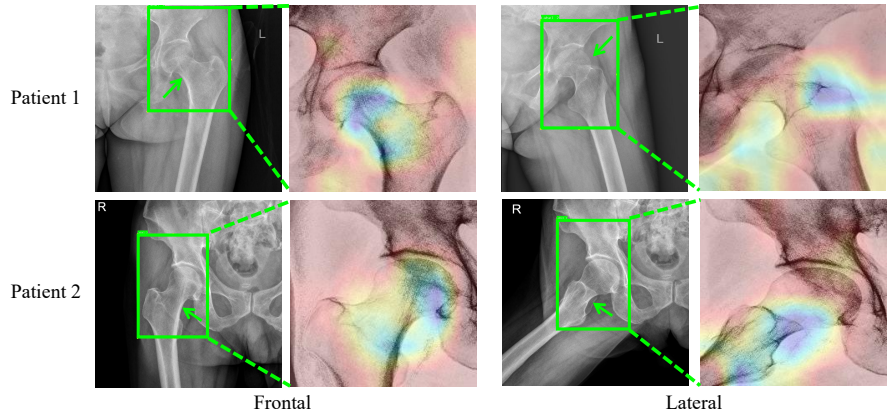


Fig. 3. Visualization results of different patients. The interest area is annotated by expert with green arrows and rectangles, whereas the highlighted areas show the interest regions of the model.

Table 1 shows the ablation results. It is observed from *Ours w/o q* and *Ours* that the proposed method improves the performance of classification by adopting the proposed cross-view deformable attention, which demonstrates that the query of Frontal-view has a positive effect on mining the discrimination features of Lateral representations. Especially, cross-view learning contributes a minimum accuracy improvement of 3% compared *Ours* to *Frontal* and *Lateral* as discriminative features between different views can be complementary. Moreover, we present the performance of different view-specific networks by comparing *Frontal* to *Frontal_swin*, the results show that the deformable transformer network gains higher accuracy with about 1.7% increment. For the Lateral-view, the deformable transformer network also has comparable performance to Swin-transformer.

Visualization results. To verify the effectiveness of the proposed framework, we visualize the interest regions of the model as shown in Fig. 3. It shows that the model could concentrate on the interested region of the diagnosis as labeled by expert. In addition, for the diagnosis of non-displaced hip fracture, the smoothness of the bone edge is a very important reference. As shown in Fig. 3, our model is also very good at focusing on bone smoothness in the same area from different perspectives in the same patient, indicating that the features from the Frontal view actually have a guidance to feature selection of the Lateral view.

4 Conclusion

This paper innovatively introduces a cross-view deformable transformer framework for non-displaced hip fracture classification from paired hip X-ray images. We adopt the deformable self-attention module to locate the interested regions of each view, while exploring feature relations among Lateral-view with the guid-

ance of Frontal-view characteristics. In addition, the proposed deformable cross-view learning method is general and has great potential to boost the performance of detecting other complicated disease. Our future work will focus on more effective training strategies and extend our framework to other cross-view medical image analysis problems.

Acknowledgement. This work was supported by the National Key Research and Development Program of China (2019YFE0113900).

References

1. Annarumma, M., Withey, S.J., Bakewell, R.J., Pesce, E., Goh, V., Montana, G.: Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology **291**(1), 196–202 (2019)
2. Bekker, A.J., Shalhon, M., Greenspan, H., Goldberger, J.: Multi-view probabilistic classification of breast microcalcifications. IEEE Transactions on medical imaging **35**(2), 645–653 (2015)
3. Bertrand, H., Hashir, M., Cohen, J.P.: Do lateral views help automated chest x-ray predictions? arXiv preprint arXiv:1904.08534 (2019)
4. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K.: Deep learning for chest x-ray analysis: A survey. Medical Image Analysis **72**, 102125 (2021)
5. Carneiro, G., Nascimento, J., Bradley, A.P.: Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions. Deep learning for medical image analysis pp. 321–339 (2017)
6. Chen, Z., Zhu, Y., Zhao, C., Hu, G., Zeng, W., Wang, J., Tang, M.: Dpt: Deformable patch-based transformer for visual recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2899–2907 (2021)
7. Cohen, J.P., Dao, L., Roth, K., Morrison, P., Bengio, Y., Abbasi, A.F., Shen, B., Mahsa, H.K., Ghassemi, M., Li, H., et al.: Predicting covid-19 pneumonia severity on chest x-ray with deep learning. Cureus **12**(7) (2020)
8. Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B.: Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134 (2022)
9. 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)
10. Hashir, M., Bertrand, H., Cohen, J.P.: Quantifying the value of lateral views in deep learning for chest x-rays. In: Medical Imaging with Deep Learning. pp. 288–303. PMLR (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
12. Huang, G., Liu, Z., Laurens, V., Weinberger, K.Q.: Densely connected convolutional networks. IEEE Computer Society (2016)
13. Krogue, J.D., Cheng, K.V., Hwang, K.M., Toogood, P., Meinberg, E.G., Geiger, E.J., Zaid, M., McGill, K.C., Patel, R., Sohn, J.H., et al.: Automatic hip fracture identification and functional subclassification with deep learning. Radiology: Artificial Intelligence **2**(2), e190023 (2020)

14. Li, M.D., Arun, N.T., Gidwani, M., Chang, K., Deng, F., Little, B.P., Mendoza, D.P., Lang, M., Lee, S.I., O'Shea, A., et al.: Automated assessment of covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. MedRxiv pp. 2020–05 (2020)
15. Liu, Y., Zhang, F., Zhang, Q., Wang, S., Wang, Y., Yu, Y.: Cross-view correspondence reasoning based on bipartite graph convolutional network for mammogram mass detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3812–3822 (2020)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. Mutasa, S., Varada, S., Goel, A., Wong, T.T., Rasiej, M.J.: Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. Journal of Digital Imaging **33**, 1209–1217 (2020)
18. Novikov, A.A., Lenis, D., Major, D., Hladvka, J., Wimmer, M., Bühler, K.: Fully convolutional architectures for multiclass segmentation in chest radiographs. IEEE transactions on medical imaging **37**(8), 1865–1876 (2018)
19. Rubin, J., Sanghavi, D., Zhao, C., Lee, K., Qadir, A., Xu-Wilson, M.: Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. arXiv preprint arXiv:1804.07839 (2018)
20. Shokouh, G.S., Magnier, B., Xu, B., Montesinos, P.: Ridge detection by image filtering techniques: a review and an objective analysis. Pattern Recognition and Image Analysis **31**, 551–570 (2021)
21. Tang, Y.X., Tang, Y.B., Peng, Y., Yan, K., Bagheri, M., Redd, B.A., Brandon, C.J., Lu, Z., Han, M., Xiao, J., et al.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ digital medicine **3**(1), 70 (2020)
22. Van Tulder, G., Tong, Y., Marchiori, E.: Multi-view analysis of unregistered medical images using cross-view transformers (2021)
23. Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4794–4803 (2022)
24. Yamada, Y., Maki, S., Kishida, S., Nagai, H., Arima, J., Yamakawa, N., Iijima, Y., Shiko, Y., Kawasaki, Y., Kotani, T., et al.: Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. Acta Orthopaedica **91**(6), 699–704 (2020)
25. Yue, X., Sun, S., Kuang, Z., Wei, M., Torr, P.H., Zhang, W., Lin, D.: Vision transformer with progressive sampling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 387–396 (2021)
26. Zhu, X., Feng, Q.: Mvc-net: Multi-view chest radiograph classification network with deep fusion. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 554–558. IEEE (2021)