Improving Pathology Localization: Multi-series Joint Attention Takes the Lead

Ashwin Raju[✉], Micha Kornreich, Colin Hansen, James Browning, Jayashri Pawar, Richard Herzog, Benjamin Odry, and Li Zhang

Covera Health, New York, NY, USA
`ashwin.raju@coverahealth.com`

Abstract. Automated magnetic resonance imaging (MRI) pathology localization can significantly reduce inter-reader variability and the time expert radiologists need to make a diagnosis. Many automated localization pipelines only operate on a single series at a time and are unable to capture inter-series relationships of pathology features. However, some pathologies require the joint consideration of multiple series to be accurately located in the face of highly anisotropic volumes and unique anatomies. To efficiently and accurately localize a pathology, we propose a **M**ulti-series j**O**int **AT**tention localization framework (MOAT) for MRI, which shares information among different MRI series to jointly predict the pathological location(s) in each MRI series. The framework allows different MRI series to share latent representations with each other allowing each series to get location guidance from the others and enforcing consistency between the predicted locations. Extensive experiments on three knee MRI pathology datasets, including medial compartment cartilage (MCC) high-grade defects, medial meniscus (MM) tear and displaced fragment/flap (DF) with 2729, 2355, and 4608 studies respectively, show that our proposed method outperforms the state of the art approaches by 3.4 to 8.0 mm on L1 distance, 6 to 27% on specificity and 5 to 14% on sensitivity across different pathologies.

Keywords: Pathology localization · Multi-series · Self-attention

1 Introduction

MRI is an essential diagnostic and investigative tool in clinical and research settings. Expert radiologists rely on multiple MRI series of varying acquisition parameters and orientations to capture different aspects of the underlying anatomy and diagnose any defect or pathology that may be present. For a knee

study, it is typical to acquire MRI series with coronal, sagittal, and axial orientations using proton density (PD), proton density fat suppressed (PDFS) or T2-weighted fat suppressed series (T2FS) for each study. When series are analyzed in concert, a radiologist can make a more effective diagnosis and mark down the location of any corresponding defect in each series. The defect location is typically represented as a single point [3] regardless of the defect size as a balance of effectiveness and efficiency.

In recent years, convolutional neural networks (CNNs) have achieved promising results in pathology localization. Many approaches rely on generating a multi-variate Gaussian heatmap, where the peak of the distribution represents the pathology localization. Hourglass [11,16], an encoder-decoder style architecture [9], is a mainstream model to generate a Gaussian heatmap. It uses a series of convolutional and pooling layers to extract features from the input image followed by upsampling and convolutional layers to generate the Gaussian heatmap. However, Hourglass-based methods can be overly resource-intensive when applied to 3D volumes [11]. To overcome this, regression-based models are becoming popular for detecting defects wherein a fully-connected layer is used on top of the encoder blocks to directly predict the location. These methods also alleviate the need for heatmap generation and post-processing methods to compute the location. Recently, transformer-based models have emerged as a promising trend in localization [4,6,14], and their performance has exceeded that of encoder-decoder based methods on single MRI volumes [4,7]. With the availability of multiple series, we propose a framework that imitates a clinical workflow, by simultaneously analyzing multiple series and paying attention to the location that corresponds to a pathology across multiple series.

To do this, we design a framework that utilizes self-attention across multiple series and we further add a mask to allow the model to focus on relevant areas, which we term as Masked Self-Attention (MSA). To predict the pathology location, we use a transformer decoder with an encoder-based initialization of the reference points. This approach provides a strong initial guess of the pathology location, improving the accuracy of the model's predictions. Overall, our framework leverages the strengths of both self-attention and encoder-decoder architectures to enhance the performance of pathology localization.

Specifically, our contributions are:

– We introduce a framework that enables the simultaneous use of multiple series from an MRI study, allowing for the sharing of pathology information across different series through Masked Self-Attention.
– We design a transformer-based decoder model to predict consistent locations across series in an MRI study, which reduces the network's parameters compared to standard heatmap-based approaches.
– Through extensive experiments on three knee pathologies, we demonstrate the effectiveness and efficiency of our framework, showing the benefits of Masked Self-Attention and a Pathology localization decoder to accurately predict pathology locations.

Overall, our framework represents a promising step towards more consistent and accurate localization, which could have important applications in medical diagnosis and treatment.
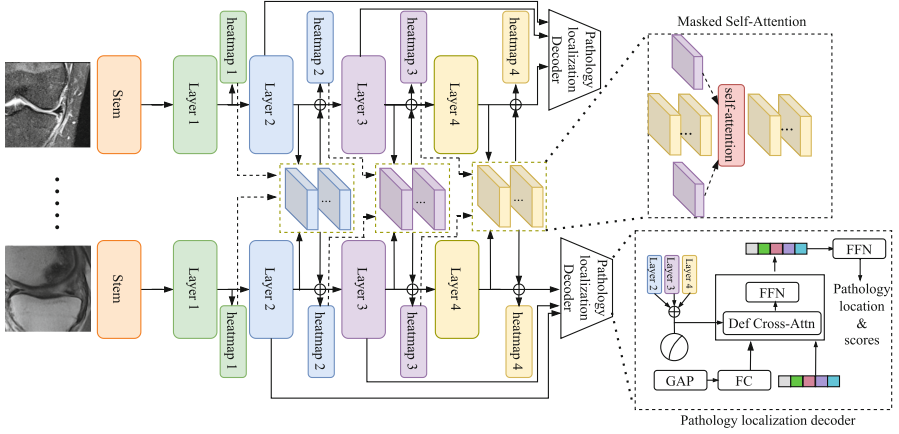

Fig. 1. Overview. More than 1 series are passed to encoders that have shared parameters. "Stem", "layer1", "layer2", "layer3" and "layer4" follows the ResNet [12] architecture convention. We perform Masked Self-Attention starting from layer 2. The Pathology localization decoder accepts feature maps from layer 2 to layer 4 and uses a query for each series to perform deformable cross attention to generate pathological landmarks.

2   Methods

2.1   Our Architecture

We aim to produce a reliable pathology location for each series in a given study if a location is available for that series. More formally, we assume that we are given a dataset, $\mathcal{D} = \{X_i, Y_i\}_{i=1}^{N}$, with N denoting the total number of studies in the dataset, X_i and Y_i denoting the set of series and corresponding location for each series. Due to different acquisition protocols, the number of series in each X_i can vary. Similarly, each Y_i can have a different number of location. Our goal is to predict a pathology location for each series and its corresponding confidence score. Figure 1 outlines our framework which can accept multiple series to generate a more accurate locations for each series.

2.2   Backbone

Our framework contains a backbone, which is responsible for generating multi-level feature maps. The multi-level feature maps are then fed into the pathology

localization decoder. We use a 3D ResNet50 [12], which accepts the volume as the input and generates multiple feature maps. Each series has its own backbone with the weights been shared. Given an input $X_i^k \in \mathbb{R}^{d \times w \times h}$, denoting a series k from the study i, we extract multiple feature maps of resolutions $F^j \in (\frac{d}{1}, \frac{w}{8}, \frac{h}{8}), (\frac{d}{2}, \frac{w}{16}, \frac{h}{16}), (\frac{d}{4}, \frac{w}{32}, \frac{h}{32})$ for each series k. We adhere to common standards by initializing the 3D ResNet50 backbone with pretrained weights. Prior work fine-tunes weights from the ImageNet dataset, but this may not be optimal if the target dataset has different characteristics. Our pretrained model for medical image analysis is based on ConVIRT [15], which uses visual representations and descriptive text from our internal dataset that contains 35433 image and text pairs.

2.3   Masked Self-attention

To explore the complementary information between different series, we use Masked Self-Attention inspired from [2] which we call MSA, a powerful tool commonly used in multi-modality [8,10] models that enable to capture long-range dependencies between features. More formally, we denote the latent feature maps $R_l = \{F_l^j\}_{j=1}^J$, where j and l represents j^{th} series and l^{th} layer, $F_l^j \in (C_{in} \times d' \times w' \times h')$ with C_{in} representing the number of channels, d' representing the depth dimension, and w' and h' representing the width and height dimensions, respectively. We concatenate the features F_l^j along the depth dimension d' and add position embedding on the concatenated features. The transformer uses a linear projection for computing the set of queries, keys and values Q, K and V respectively. We adhere to the naming conventions used in [8].

$$Q = R_l.U^q, K = R_l.U^k, V = R_l.U^v \qquad (1)$$

where $U^q \in \mathbb{R}^{C_{in} \times C_q}$, $U^k \in \mathbb{R}^{C_{in} \times C_k}$ and $U^v \in \mathbb{R}^{C_{in} \times C_v}$. The self-attention is calculated by taking the dot products between Q and K and then aggregating the values for each query,

$$A = \text{Softmax}\left(M_{l-1} + B + \frac{QK^T}{\sqrt{C_k}}\right) V \qquad (2)$$

$$M_{l-1} = \begin{cases} 0 & \text{if } M_{l-1} = 1 \\ -\infty & \text{otherwise} \end{cases} \qquad (3)$$

where, the attention mask $M_{l-1} \in \{0,1\}$ is a binarized output (thresholded at δ_t) of the the resized mask prediction of the previous $(l-1)$-th layer. δ_t is empirically set to 0.15. The attention mask ignores the features that are not relevant to the pathology and attends to pathological features. B is a mask to handle missing series and it shares the same equation as 3.

Finally, the transformer uses a non-linear transformation to calculate the output features, R_{l+1}, which shares the same resolution as that of R_l.

$$R_{l+1} = \text{MLP}(A) + R_l \qquad (4)$$

The transformer applies the attention mechanism 3 L times to generate a deep representation learning among the features. This approach allows the transformer model to effectively capture the relationships between different input positions.

2.4   Pathology Localization Decoder

The localization decoder follows the transformer decoder paradigm, using a query, reference points, and input feature maps to predict a location and corresponding score. The decoder has N identical layers, each consisting of cross-attention and feed forward networks (FFNs). The query $Q \in \mathbb{R}^{1 \times 256}$ and reference points $R \in \mathbb{R}^3$ go through each layer, generating an updated Q as input for the next layer. Unlike Deformable DETR [17], the decoder initializes reference points by taking the last layer of the backbone feature map and applying Global Average Pooling, followed by a fully connected layer to generate the initial reference point. The localization refinement stage outputs location and scores for each layer N_i, similar to Deformable DETR, providing fast convergence.

2.5   Loss Functions

The model generates a single location $\hat{y}_l \in \mathbb{R}^3$, score y_s and auxiliary heatmap outputs H for each series in a given study. The goal of our framework is to generate one reasonable location and its corresponding score for each series. Since there may be multiple locations annotated for a series, we use the Hungarian Matching function [17] to find optimal matching with the prediction to one of many ground truth locations. This is similar to the approach used in DETR. The Masked Self-Attention in our framework uses heatmaps generated from the previous layers. To ensure accurate heatmap generation, we apply an auxiliary heatmap loss using Mean Square Error (MSE) between the generated heatmap and the ground truth Gaussian heatmap, where the loss is defined as,

$$L_{heatmap} = \sum_{i=1}^{K}(x - h_i)^2 \tag{5}$$

where K is the number of intermediate heatmaps generated, x and h_i are ground truth heatmap and predicted heatmap. To penalize the predicted location, we use the Huber loss defined as,

$$L_{point} = \sum_{i=1}^{N} \begin{cases} \frac{1}{2}(y_l^i - \hat{y_l^i})^2 & \text{if } \left|(y_l^i - \hat{y_l^i})\right| < \delta \\ \delta((y_l^i - \hat{y_l^i}) - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{6}$$

where δ is empirically set to 0.3. The distance of a pathology does not differ more than λ (which can be calculated from the dataset) across series. With this information, we enforce proximity between the world coordinates which can be converted from the predicted volume coordinates across different series. We

employ a Margin L1 loss, which penalizes the distance between points if they exceed the margin. Formally,

$$L_{cons} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} max(0, L1(wc_{\hat{y}_l^i}, wc_{\hat{y}_l^j}) - \lambda) \tag{7}$$

where N is the number of series in a given study, $wc_{\hat{y}_l}$ is the world coordinates converted from volume coordinates.

We then formulate the confidence score loss by considering the sum over the series of the binary cross entropy between the ground truth confidence score and predicted confidence score, formally defined as,

$$L_{score} = \sum_{i=1}^{N} -(y_s^i \log(p_i) + (1 - y_s^i) \log(1 - p_i)) \tag{8}$$

Overall, the entire loss for a given study is formulated as,

$$L = w_1 L_{point} + w_2 L_{score} + w_3 L_{cons} + w_4 L_{heatmap} \tag{9}$$

We set the hyper parameter w_1, w_2, w_3 and w_4 as 10, 1, 0.1, 1 respectively. These values are empirically set based on the validation loss.

3   Experiment

3.1   Implementation Details, Datasets and Evaluation Protocols

Implementation Details. Our model was implemented in Pytorch 1.13.1 on a NVIDIA A6000 GPU. We used an AdamW [5] optimizer with a weight decay of 10^{-4}. The initial learning rate for encoder was empirically set as 10^{-5} and 10^{-4} for all other modules. Before running the pathology detection, we perform a pre-processing step similar to [3] and resize the volume to $28 \times 128 \times 128$. Furthermore, we clip the intensity of the images at the 1st and 99th percentile, followed by an intensity normalization to ensure a mean of 0 and standard deviation of 1. Other hyper-parameters are mentioned in the supplementary paper.

Datasets. The study is limited to secondary use of existing HIPPA-based de-identified data. No IRB required. We primarily conduct our experiments using knee MRI datasets, with a specific focus on MM tear, MM displaced fragment flap (DF), and MCC defect. Studies were collected at over 25 different institutions, and differed in scanner manufacturers, magnetic field strengths, and imaging protocols. The pathological locations were annotated by American Board certified sub-specialists radiologists. The most common series types included fat-suppressed (FS) sagittal (Sag), coronal (Cor) and axial (Ax) orientations, using either T2-weighted (T2) or proton-density (PD) protocols. For pathology detection, we use CorFS, SagFS, and SagPD. The dataset statistics that we use for training, validation and test are shown in Table 1.

Table 1. Cor, Sag, SagPD refers to Coronal FS, Sagittal FS, Sagittal PD respectively. Values under series refer to number of series where defect locations are available. Negatives refer to number of studies that does not have a pathology.

Pathology	Train					Validation				Test			
	Negatives	Cor	Sag	SagPD	Union	Cor	Sag	SagPD	Union	Cor	Sag	SagPD	Union
MM Tear	1215	1466	1173	1146	**2679**	975	843	771	**975**	954	857	765	**954**
DF	673	387	306	243	**1862**	247	176	148	**277**	190	146	106	**216**
MCC	1000	759	797	364	**1926**	360	387	152	**437**	304	317	136	**366**

Evaluation Protocols. A useful pathology detection device should point the user to the correct location of a pathology. For model evaluation, we use the L1 distance between the predicted location to any annotation of the same pathology, labeled on the same series. To evaluate the pathology localization in a given study, we use the predicted pathology localization mask, which is obtained by thresholding the confidence score.

However, this alone does not provide a complete picture of the model's performance. To evaluate our confidence score's performance, we analyze the specificity and sensitivity of the confidence scores. We report the mean over all series in the test studies in Table 2

Table 2. Quantitative results. We show the L1 distance measured in (mm), Sensitivity (Sn), and Specificity (Sp) score for different models. "*" refers to the models that were trained with different hyper-parameters from their mentioned ones. The results are evaluated on the test dataset.

Methods	Param	FLOPs	MM Tear			MM DF			MCC defect		
	(M)	(G)	L1↓	Sn↑	Sp↑	L1↓	Sn↑	Sp↑	L1↓	Sn ↑	Sp↑
UNet* [3]	54.6	117.3	10.1	0.63	0.71	16.1	0.53	0.62	9.5	0.70	0.71
UNet w MSA	63.7	135.4	9.4	0.70	0.71	15.3	0.59	0.68	8.1	0.70	0.72
KNEEL* [11]	74.2	152.1	9.1	0.68	0.70	14.5	0.61	0.70	9.1	0.69	0.72
Regression	24.1	45.7	17.2	0.71	0.75	21.4	0.70	0.72	10.1	0.72	0.74
DETR* [1]	51.4	67.3	14.6	0.75	0.73	20.2	0.72	0.71	13.4	0.72	0.74
Def. DETR* [17]	37.1	71.4	15.8	0.77	0.80	21.3	0.73	0.75	12.9	0.75	0.79
Poseur* [6]	35.1	65.1	13.1	0.76	0.80	17.3	0.71	0.75	11.5	0.78	0.79
Poseur w MSA	44.2	71.3	10.3	0.80	0.81	14.9	0.73	0.76	7.2	0.80	**0.86**
MOAT	28.2	63.4	**4.7**	**0.85**	**0.86**	**8.1**	**0.80**	**0.81**	**3.9**	**0.88**	**0.86**

3.2 Comparison with SOTA Methods

Heatmap-Based Architectures. The proposed architecture was compared to two other models, the Gaussian Ball approach [3] which utilizes a UNet architecture to generate a heatmap and KNEEL [11], which uses an hourglass network architecture to predict the Gaussian heatmap. Two variants of UNet were compared, one with MSA and one without. The threshold was set for each model

which balanced sensitivity and specificity on the validation data. The comparison revealed that the sensitivity and specificity of the proposed MOAT model were 14 to 27% and 15 to 17% higher, respectively, than those of the other models. Additionally, the L1 distance of the heatmap-based model was approximately 5.4 to 8.0 mm higher than that of MOAT for all true positives. Overall, the results suggest that MOAT outperforms the other models in terms of sensitivity, specificity, and L1 distance.

Regression-Based Architectures. We compared our proposed architecture with several other methods: 1) a simple regression method that removes the pathology localization decoder and uses a fully connected layer to predict the pathology locations, 2) DETR, 3) deformable DETR [17], and 4) Poseur [6], which uses Residual Log estimation. We adopt our ConVIRT pretrained encoder and add MSA to all the regression models to ensure a fair comparison. MOAT, which has 63.4G FLOPs, is highly efficient when compared to State-Of-The-Art (SOTA) regression models and has L1 distance lower than other models (4.7 mm) and the highest sensitivity and specificity among the models. We attach the standard deviation scores for each model in the supplementary section.

3.3   Ablation Study

We first analyze the importance of MSA to our framework by training models with and without MSA. As MSA is a variant of self-attention, we also experiment with self-attention and with an attention mechanism [13] that was popular prior to self-attention. Table 4 shows the L1 distance for Medial Meniscus Tear (MM Tear) pathology, where our MSA which is a variant of self-attention is able to achieve the lowest L1 distance. Similarly, we analyze the weight factor for consistency loss, as different weight factor yields different results. From Table 3, we can see that the lowest L1 distance was obtained when the weight factor was 0.1. All the ablation studies were performed on the MM Tear validation dataset.

Table 3. Ablation study on MM Tear dataset to analyze the need for Masked Self-Attention.

Methods	L1 distance ↓
No Masked Self-Attention	11.3
self-attention	8.8
Masked Self-Attention	**6.1**
CBAM [13]	16.2

Table 4. Ablation study on MM Tear to analyze the weight factor for the consistency loss.

Consistency(w_3)	L1 distance ↓
10	11.2
1	6.4
0.1	6.1
0.01	**5.1**

4   Conclusion

We propose MOAT, a framework for performing localization in multi-series MRI studies which benefits from the ability to share relevant information across series

via a novel application of self-attention. We increase the efficiency of the MOAT model by using a pathology localization decoder which is a variant of deformable decoder and initializes the reference points from the backbone of the model. We evaluate the effectiveness of our proposed framework (MOAT) on three challenging pathologies from knee MRI and find that it represents a significant improvement over several SOTA localization techniques. Moving forward, we aim to apply our framework to pathologies from other body parts with multiple series.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation (2022)
3. Kornreich, M., et al.: Combining mixed-format labels for AI-based pathology detection pipeline in a large-scale knee MRI study. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 183–192. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_18
4. Li, X., et al.: SDMT: spatial dependence multi-task transformer network for 3D knee MRI segmentation and landmark localization. IEEE Trans. Med. Imaging (2023)
5. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
6. Mao, W., et al.: Poseur: direct human pose regression with transformers (2022)
7. Mathai, T.S., et al.: Lymph node detection in T2 MRI with transformers. In: Medical Imaging 2022: Computer-Aided Diagnosis, vol. 12033, pp. 855–859. SPIE (2022)
8. Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Shvetsova, N., et al.: Everything at once - multi-modal fusion transformer for video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20020–20029 (2022)
11. Tiulpin, A., Melekhov, I., Saarakkala, S.: Kneel: knee anatomical landmark localization using hourglass networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
12. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. CoRR abs/1711.11248 (2017). http://arxiv.org/abs/1711.11248
13. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

14. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose+: vision transformer foundation model for generic body pose estimation. arXiv preprint arXiv:2212.04246 (2022)
15. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference, pp. 2–25. PMLR (2022)
16. Zhu, J., Zhao, Q., Zhu, J., Zhou, A., Shao, H.: A novel method for 3D knee anatomical landmark localization by combining global and local features. Mach. Vis. Appl. **33**(4), 52 (2022)
17. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)