MRIS: A Multi-modal Retrieval Approach for Image Synthesis on Diverse Modalities

Boqi Chen[(✉)] and Marc Niethammer

Department of Computer Science, University of North Carolina at Chapel Hill,
Chapel Hill, USA
`bqchen@cs.unc.edu`

Abstract. Multiple imaging modalities are often used for disease diagnosis, prediction, or population-based analyses. However, not all modalities might be available due to cost, different study designs, or changes in imaging technology. If the differences between the types of imaging are small, data harmonization approaches can be used; for larger changes, direct image synthesis approaches have been explored. In this paper, we develop an approach, MRIS, based on multi-modal metric learning to synthesize images of diverse modalities. We use metric learning via multi-modal image retrieval, resulting in embeddings that can relate images of different modalities. Given a large image database, the learned image embeddings allow us to use k-nearest neighbor (k-NN) regression for image synthesis. Our driving medical problem is knee osteoarthritis (KOA), but our developed method is general after proper image alignment. We test our approach by synthesizing cartilage thickness maps obtained from 3D magnetic resonance (MR) images using 2D radiographs. Our experiments show that the proposed method outperforms direct image synthesis and that the synthesized thickness maps retain information relevant to downstream tasks such as progression prediction and Kellgren-Lawrence grading (KLG). Our results suggest that retrieval approaches can be used to obtain high-quality and meaningful image synthesis results given large image databases. Our code is available at https://github.com/uncbiag/MRIS.

Keywords: Metric learning · k-nearest neighbor · osteoarthritis

1 Introduction

Recent successes of machine learning algorithms in computer vision and natural language processing suggest that training on large datasets is beneficial for model performance [2,5,18,21]. While several efforts to collect very large medical image datasets are underway [12,19], collecting large *homogeneous* medical

image datasets is hampered by: a) cost, b) advancement of technology throughout long study periods, and c) general heterogeneity of acquired images across studies, making it difficult to utilize all data. Developing methods accounting for different imaging types would help make the best use of available data.

Although image harmonization and synthesis [3,14,15,22] methods have been explored to bridge the gap between different types of imaging, these methods are often applied to images of the same geometry. On the contrary, many studies acquire significantly more diverse images; e.g., the OAI image dataset[1] [9] contains both 3D MR images of different sequences and 2D radiographs. Similarly, the UK Biobank [19] provides different 3D MR image acquisitions and 2D DXA images. Ideally, a machine learning system can make use of all data that is available. As a related first step in this direction, we explore the feasibility of predicting information gleaned from 3D geometry using 2D projection images. Being able to do so would allow a) pooling datasets that drastically differ in image types or b) relating information from a cheaper 2D screening to more readily interpretable 3D quantities that are difficult for a human observer.

We propose an image synthesis method for diverse modalities based on multi-modal metric learning and k-NN regression. To learn the metric, we use image retrieval as the target task, which aims at embedding images such that matching pairs of different modalities are close in the embedding space. We use a triplet loss [24] to contrastively optimize the gap between positive and negative pairs based on the cosine distance over the learned deep features. In contrast to the typical learning process, we carefully design the training scheme to avoid interference when training with longitudinal image data. Given the learned embedding, we can synthesize images between diverse image types by k-NN regression through a weighted average based on their distances measured in the embedding space. Given a large database, this strategy allows for a quick and simple estimation of one image type from another.

We use knee osteoarthritis as the driving medical problem and evaluate our proposed approach using the OAI image data. Specifically, we predict cartilage thickness maps obtained from 3D MR images using 2D radiographs. This is a highly challenging task and therefore is a good test case for our approach for the following reasons: 1) cartilage is not explicitly visible on radiographs. Instead, the assessment is commonly based on joint space width (JSW), where decreases in JSW suggest decreases in cartilage thickness [1]; 2) the difficulty in predicting information obtained from a 3D image using only the 2D projection data; 3) the large appearance difference between MR images and thickness maps; 4) the need to capture fine-grained details within a small region of the input radiograph. While direct regression via deep neural networks is possible, such approaches lack interpretability and we show that they can be less accurate for diverse images.

The main contributions of our work are as follows.

1. We propose an image synthesis method for diverse modalities based on multi-modal metric learning using image retrieval and k-NN regression. We carefully construct the learning scheme to account for longitudinal data.

---

[1] https://nda.nih.gov/oai/.

2. We extensively test our approach for osteoarthritis, where we synthesize cartilage thickness maps derived from 3D MR using 2D radiographs.
3. Experimental results show the superiority of our approach over commonly used image synthesis methods, and the synthesized images retain sufficient information for downstream tasks of KL grading and progression prediction.
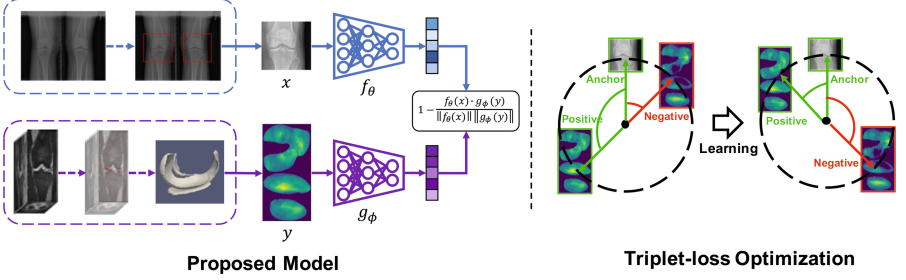
Fig. 1. Proposed multi-modal metric learning model (left) trained using a triplet loss (right). Left top: encoding the region of interest from radiographs, extracted using the method from [26]. Left bottom: encoding thickness maps, extracted from MR images using the method from [11]. Features are compared using cosine similarity. Right: applying triplet loss on cosine similarity, where nonpaired data is moved away from paired data.

2 Method

In this work, we use multi-modal metric learning followed by k-NN regression to synthesize images of diverse modalities. Our method requires 1) a database containing matched image pairs; 2) target images aligned to an atlas space.

2.1 Multi-modal Longitudinally-Aware Metric Learning

Let $\{(x_a^i, y_a^i)\}$ be a database of multiple paired images with each pair containing two modalities x and y of the a-th subject and i-th timepoint if longitudinal data is available. We aim to learn a metric that allows us to reliably identify related image pairs, which in turn relate structures of different modalities. Specifically, we train our deep neural network via a triplet loss so that matching image pairs are encouraged to obtain embedding vectors closer to each other than mismatched pairs. Figure 1 illustrates the proposed multi-modal metric learning approach, which uses two convolutional neural networks (CNNs), each for extracting the features of one modality. The two networks may share the same architecture, but unlike Siamese networks [4], our CNNs have independent sets of weights. This is because the two modalities differ strongly in appearance.

Denoting the two CNNs as $f(\cdot; \theta)$ and $g(\cdot; \phi)$, where θ and ϕ are the CNN parameters, we measure the feature distance between two images x and y using cosine similarity

$$d(x,y) = 1 - \frac{f(x;\theta) \cdot g(y;\phi)}{\|f(x;\theta)\| \, \|g(y;\phi)\|} \,, \tag{1}$$

where the output of f and g are vectors of the same dimension[2]. Given a mini-batch of N paired images, our goal is to learn a metric such that $f(x_a^i)$ and $g(y_a^i)$ are close (that is, for the truly matching image pair), while $f(x_a^i)$ and $g(y_b^j)$ are further apart, where $a \neq b$ and i, j are arbitrary timepoints of subjects a, b, respectively. *We explicitly avoid comparing across timepoints of the same subject to avoid biasing longitudinal trends.* This is because different patients have different disease progression speeds. For those with little to no progression, images may look very similar across timepoints and should therefore result in similar embeddings. It would be undesirable to view them as negative pairs. Therefore, our multi-modal longitudinally-aware triplet loss becomes

$$loss(\{(x_a^i, y_a^i)\}) = \sum_{(a,i)} \sum_{(b,j),b \neq a} \max[d(f_\theta(x_a^i), g_\phi(y_a^i)) - d(f_\theta(x_a^i), g_\phi(y_b^j)) + m, 0] \,,$$
$$\tag{2}$$

where m is the margin for controlling the minimum distance between positive and negative pairs. We sum over all subjects at all timepoints for each batch.

To avoid explicitly tracking the subjects in a batch, we can simplify the above equation by randomly picking one timepoint per subject during each training epoch. This then simplifies our multi-modal longitudinally aware triplet loss to a standard triplet loss of the form

$$loss(\{(x_a, y_b)\}) = \sum_{a=1}^{N} \sum_{b=1,b \neq a}^{N} \max[d(f_\theta(x_a), g_\phi(y_a)) - d(f_\theta(x_a), g_\phi(y_b)) + m, 0] \,.$$
$$\tag{3}$$

2.2   Image Synthesis

After learning the embedding space, it can be used to find the most relevant images with a new input, as shown in Fig. 2. Specifically, the features of a query image x are first extracted by the CNN model f_θ we described previously. Given a database of images of the target modality $\mathcal{S}^I = \{y_a^i\}$ and their respective embeddings $\mathcal{S}^F = \{g(y_a^i)\}$, we can then select the top k images with the smallest cosine distance, which will be the most similar images given this embedding. Denoting these k most similar images as $\mathcal{K} = \{\tilde{y}^k\}$ we can synthesize an image, $\hat{y}$ based on a query image, x as a weighted average of the form

$$\hat{y} = \sum_{i=1}^{K} w_i \tilde{y}^i \quad where \quad w_i = \frac{1 - d(x, \tilde{y}^i)}{\sum_{j=1}^{K}(1 - d(x, \tilde{y}^j))} \,, \tag{4}$$

---

[2] For notational clarity we will suppress the dependency of f on θ and will write $f_\theta(\cdot)$ instead of $f(\cdot; \theta)$.

where the weights are normalized weights based on the cosine similarities. This requires us to work in an atlas space for the modality y, where all images in the database $\mathcal{S}^I$ are spatially aligned. However, images of the modality x do not need to be spatially aligned, as long as sensible embeddings can be captured by f_θ. As we will see, this is particularly convenient for our experimental setup, where the modality x is a 2D radiograph and the modality y is a cartilage thickness map derived from a 3D MR image, which can easily be brought into a common atlas space. As our synthesized image, $\hat{y}$, is a weighted average of multiple spatially aligned images, it will be smoother than a typical image of the target modality. However, we show in Sect. 3 that the synthesized images still retain the general disease patterns and retain predictive power.

Note also that our goal is not image retrieval or image reidentification, where one wants to find a known image in a database. Instead, we want to synthesize an image for a patient who is not included in our image database. Hence, we expect that no perfectly matched image exists in the database and therefore set $k > 1$. Based on theoretical analyses of k-NN regression [6], we expect the regression results to improve for larger image databases.

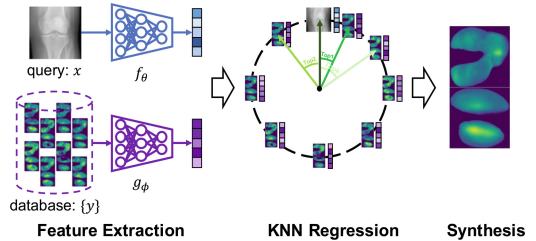## 3   Experimental Results

This section focuses on investigating the following questions on the OAI dataset:

1. *How good is our retrieval performance?* We calculate recall values to determine the performance to retrieve the correct image;

2. *How accurate are our estimated images?* We compare the predicted cartilage thickness maps with those obtained from 3D MR images;

Fig. 2. Image synthesis by k-NN regression from the database. Given an unseen image x, we extract its features $f_\theta(x)$, find the k nearest neighbors in the database $\{y\}$ based on these features, and use them for a weighted k-NN regression.

3. *Does our prediction retain disease-relevant information for downstream tasks?* We test the performance of our predicted cartilage thickness maps in predicting KLG and osteoarthritis progressors;

4. *How does our approach compare to existing image synthesis models?* We show that our approach based on simple k-NN regression compares favorably to direct image synthesis approaches.

3.1   Dataset

We perform a large-scale validation of our method using the Osteoarthritis Initiative (OAI) dataset on almost 40,000 image pairs. This dataset includes $4,796$

patients between the ages of 45 to 79 years at the time of recruitment. Each patient is longitudinally followed for up to 96 months.

Images. The OAI acquired images of multiple modalities, including T2 and DESS MR images, as well as radiographs. We use the paired DESS MR images and radiographs in our experiments. After excluding all timepoints when patients do not have complete MR/radiograph pairs, we split the dataset into three sets by patient (i.e., data from the same patient are in the same sets): Set 1) to train the image retrieval model (2,000 patients; 13,616 pairs). This set also acts as a database during image synthesis; Set 2) to train the downstream task (1,750 patients; 16,802 pairs); Set 3) to test performance (897 patients; 8,418 pairs).
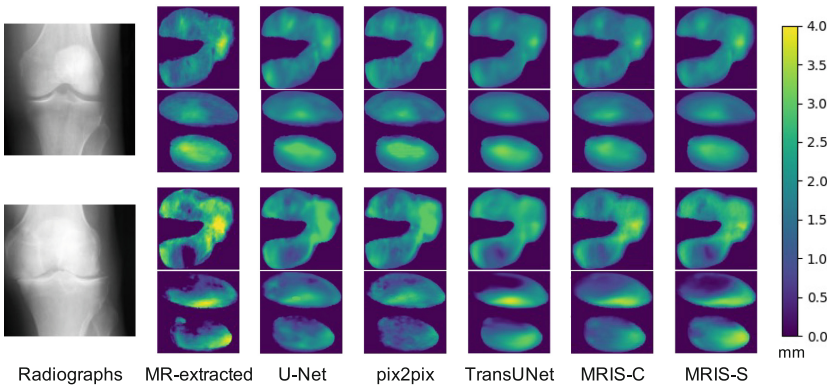
Fig. 3. Thickness map predictions for different methods and different severity. Our approach shows a better match of cartilage thickness with the MR-extracted thickness map than the other approaches. See more examples in the appendix.

Preprocessing. As can be seen from the purple dashed box in Fig. 1, we extract cartilage thickness maps from the DESS MR images using a deep segmentation network [27], register them to a common 3D atlas space [25], and then represent them in a common flattened 2D atlas space [11]. These 2D cartilage thickness maps are our target modality, which we want to predict from the 2D radiographs. Unlike MR images for which a separate scan is obtained for the left and right knees, OAI radiographs include both knees and large areas of the femur and tibia. To separate them, we apply the method proposed in [26], which automatically detects keypoints between the knee joint. As shown in the blue dashed box in Fig. 1, the region of interest for each side of the knee is being extracted using a region of 140 mm * 140 mm around the keypoints.

We normalize all input radiographs by linearly scaling the intensities so that the smallest 99% values are mapped to $[0, 0.99]$. We horizontally flip all right knees to the left as done in [11], randomly rotate images up to 15 degrees, add Gaussian noise, and adjust contrast. Unlike the radiographs, we normalize the cartilage thickness map by dividing all values by 3, which is approximately the 95-th percentile of cartilage thickness. All images are resized to $256 * 256$.

Table 1. Thickness map retrieval recall percentage on the testing set. R@k shows the percentage of queries for which the correct one is retrieved within the top k nearest neighbors.

Method	R@1 ↑	R@5 ↑	R@10 ↑	R@20 ↑
Femoral	28.26	58.19	71.13	82.11
Tibial	30.49	61.48	73.36	83.33
Combined	**45.21**	**75.53**	**84.73**	**90.64**

Table 2. Median $\pm$ MAD absolute error for both femoral and tibial cartilage between the predicted thickness maps and those extracted from MR images. We stratify the result by KLG. Larger KLG results in less accurate synthesis.

Median $\pm$ MAD		KLG01 ↓	KLG2 ↓	KLG3 ↓	KLG4 ↓	All ↓
Femoral Cartilage	U-Net	0.288 ± 0.173	0.324 ± 0.195	0.358 ± 0.214	0.410 ± 0.252	0.304 ± 0.183
	pix2pix	0.289 ± 0.173	0.326 ± 0.196	0.360 ± 0.216	0.411 ± 0.253	0.306 ± 0.183
	TransUNet	0.260 ± 0.157	0.300 ± 0.180	0.326 ± 0.195	0.384 ± 0.235	0.277 ± 0.167
	MRIS-C	0.265 ± 0.158	0.298 ± 0.178	0.319 ± 0.191	0.377 ± 0.226	0.279 ± 0.167
	MRIS-S	$\mathbf{0.259 \pm 0.155}$	$\mathbf{0.295 \pm 0.176}$	$\mathbf{0.319 \pm 0.191}$	$\mathbf{0.373 \pm 0.223}$	$\mathbf{0.275 \pm 0.164}$
Tibial Cartilage	U-Net	0.304 ± 0.181	0.324 ± 0.193	0.364 ± 0.216	0.428 ± 0.270	0.316 ± 0.188
	pix2pix	0.306 ± 0.182	0.325 ± 0.194	0.367 ± 0.219	0.433 ± 0.272	0.319 ± 0.190
	TransUNet	0.269 ± 0.160	0.288 ± 0.172	0.325 ± 0.192	$\mathbf{0.371 \pm 0.254}$	0.281 ± 0.167
	MRIS-C	0.271 ± 0.160	0.291 ± 0.171	0.319 ± 0.188	0.385 ± 0.225	0.282 ± 0.166
	MRIS-S	$\mathbf{0.265 \pm 0.157}$	$\mathbf{0.283 \pm 0.168}$	$\mathbf{0.313 \pm 0.187}$	0.379 ± 0.226	$\mathbf{0.276 \pm 0.163}$

3.2 Network Training

During multi-modal metric learning, our two branches use the ResNet-18 [10] model with initial parameters obtained by ImageNet pre-training [8]. We fine-tune the networks using AdamW [20] with initial learning rate 10^{-4} for radiographs and 10^{-5} for the thickness maps. The output embedding dimensions of both networks are 512. We train the networks with a batch size of 64 for a total of 450 epochs with a learning rate decay of 80% for every 150 epochs. We set the margin $m = 0.1$ in all our experiments.

For both downstream tasks, we fine-tune our model on a ResNet-18 pre-trained network with the number of classes set to 4 for KLG prediction and 2 for progression prediction. Both tasks are trained with AdamW for 30 epochs, batch size 64, and learning rate decay by 80% for every 10 epochs. The initial learning rate is set to 10^{-5} for KLG prediction and 10^{-4} for progression prediction.

3.3   Results

This section shows our results for image retrieval, synthesis, and downstream tasks based on the questions posed above. All images synthesized from MRIS are based on the weighted average of the retrieved top $k = 20$ thickness maps.

Image Retrieval. To show the importance of the learned embedding space, we perform image retrieval on the test set, where our goal is to correctly find the corresponding matching pair. Since our training process does not compare images of the same patient at different timepoints, we test using only the baseline images for each patient ($1,794$ pairs). During training, we created two thickness map variants: 1) combining the femoral and tibial cartilage thickness maps (Combined); 2) separating the femoral and tibial thickness maps (Femoral/Tibial), which requires training two networks. Table 1 shows the image retrieval recall, where R@k represents the percentage of radiographs for which the correct thickness map is retrieved within the k-nearest neighbors in the embedding space. Combined achieves better results than retrieving femoral and tibial cartilage separately. This may be because more discriminative features can be extracted when both cartilages are provided, which simplifies the retrieval task. In addition, tibial cartilage appears to be easier to retrieve than femoral cartilage.

Table 3. Results on the downstream tasks of KLG and progression prediction. Our synthesis methods overall perform better than other synthesis methods and obtain a comparable result with the MR-extracted thickness maps.

Method	KLG Prediction (accuracy) ↑					Progression Prediction	
	KLG01	KLG2	KLG3	KLG4	overall	average precision ↑	roc auc ↑
U-Net	0.819	0.321	0.778	0.545	0.719	0.242	0.606
pix2pix	0.805	0.396	0.735	0.654	0.722	0.225	0.625
TransUNet	0.797	**0.528**	<u>0.763</u>	**0.865**	0.746	0.286	0.654
MRIS-C	<u>0.865</u>	0.469	0.757	0.673	<u>0.781</u>	<u>0.299</u>	<u>0.713</u>
MRIS-S	**0.869**	0.479	**0.786**	0.718	**0.789**	**0.307**	0.702
MR-extracted	0.842	<u>0.523</u>	0.727	<u>0.795</u>	0.775	0.286	**0.739**

Image Synthesis. To directly measure the performance of our synthesized images on the testing dataset, we show the median $\pm$ MAD (median absolute deviation) absolute error compared to the thickness map extracted by MR in Table 2. We created two variants by combining or separating the femoral and tibial cartilage, corresponding to MRIS-C(ombined) and MRIS-S(eparate). Unlike the image retrieval recall results, MRIS-S performs better than MRIS-C (last column of Table 2). This is likely because it should be beneficial to mix and match separate predictions for synthesizing femoral and tibial cartilage. Moreover, MRIS-S outperforms all baseline image synthesis methods [7,13,23].

   Osteoarthritis is commonly assessed via Kellgren-Lawrence grade [16] on radiographs by assessing joint space width and the presence of osteophytes.

KLG=0 represents a healthy knee, while KLG=4 represents severe osteoarthritis. KLG=0 and 1 are often combined because knee OA is considered definitive only when KLG$\geq$ 2 [17]. To assess prediction errors by OA severity, we stratify our results in Table 2 by KLG. Both variants of our approach perform well, outperforming the simpler pix2pix and U-Net baselines for all KLG. The TransUNet approach shows competitive performance, but overall our MRIS-S achieves better results regardless of our much smaller model size. Figure 3 shows examples of images synthesized for the different methods for different severity of OA.

Downstream Tasks. The ultimate question is whether the synthesized images can still retain information for downstream tasks. Therefore, we test the ability to predict KLG and OA progression, where we define OA progression as whether or not the KLG will increase within the next 72 months. Table 3 shows that our synthesized thickness maps perform on par with the MR-extracted thickness maps for progression prediction and we even outperform on predicting KLG. MRIS overall performs better than U-Net [23], pix2pix [13] and TransUNet [7].

4   Conclusion

In this work, we proposed an image synthesis method using metric learning via multi-modal image retrieval and k-NN regression. We extensively validated our approach using the large OAI dataset and compared it with direct synthesis approaches. We showed that our method, while conceptually simple, can effectively synthesize alignable images of diverse modalities. More importantly, our results on the downstream tasks showed that our approach retains disease-relevant information and outperforms approaches based on direct image regression. Potential shortcomings of our approach are that the synthesized images tend to be smoothed due to the weight averaging and that spatially aligned images are required for the modality to be synthesized.

References

1. Altman, R.D., et al.: Radiographic assessment of progression in osteoarthritis. Arthritis Rheumatism: Official J. Am. College Rheumatol. **30**(11), 1214–1225 (1987)

2. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)

3. Boulanger, M., et al.: Deep learning methods to generate synthetic CT from MRI in radiotherapy: a literature review. Physica Med. **89**, 265–281 (2021)

4. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a Siamese time delay neural network. In: Advances in Neural Information Processing Systems 6 (1993)

5. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)

6. Chen, G.H., Shah, D., et al.: Explaining the success of nearest neighbor methods in prediction. Foundat. Trends Mach. Learn. **10**(5–6), 337–588 (2018)

7. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. Ieee (2009)

9. Eckstein, F., Wirth, W., Nevitt, M.C.: Recent advances in osteoarthritis imaging-the osteoarthritis initiative. Nat. Rev. Rheumatol. **8**(10), 622–630 (2012)

10. 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)

11. Huang, C., et al.: DADP: dynamic abnormality detection and progression for longitudinal knee magnetic resonance images from the osteoarthritis initiative. In: Medical Image Analysis, p. 102343 (2022)

12. Ikram, M.A., et al.: Objectives, design and main findings until 2020 from the Rotterdam study. Eur. J. Epidemiol. **35**(5), 483–517 (2020)

13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)

14. 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: Deeba, F., Johnson, P., Würfl, T., Ye, J.C. (eds.) MLMIR 2020. LNCS, vol. 12450, pp. 123–133. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61598-7_12

15. Kawahara, D., Nagata, Y.: T1-weighted and T2-weighted MRI image synthesis with convolutional generative adversarial networks. Reports Practical Oncol. Radiotherapy **26**(1), 35–42 (2021)

16. Kellgren, J.H., Lawrence, J.: Radiological assessment of osteo-arthrosis. Ann. Rheum. Dis. **16**(4), 494 (1957)

17. Kohn, M.D., Sassoon, A.A., Fernando, N.D.: Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin. Orthop. Relat. Res. **474**(8), 1886–1893 (2016)

18. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

19. Littlejohns, T.J., Sudlow, C., Allen, N.E., Collins, R.: UK Biobank: opportunities for cardiovascular research. Eur. Heart J. **40**(14), 1158–1166 (2019)

20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

21. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)

22. Ren, M., Dey, N., Fishbaugh, J., Gerig, G.: Segmentation-renormalized deep feature modulation for unpaired image harmonization. IEEE Trans. Med. Imaging **40**(6), 1519–1530 (2021)

23. 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

24. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)

25. Shen, Z., Han, X., Xu, Z., Niethammer, M.: Networks for joint affine and non-parametric image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4224–4233 (2019)

26. 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)

27. Xu, Z., Shen, Z., Niethammer, M.: Contextual additive networks to efficiently boost 3d image segmentations. In: DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 92–100. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_11