KNEEL: Knee Anatomical Landmark Localization Using Hourglass Networks

Aleksei Tiulpin[1,2] Iaroslav Melekhov[3]
Simo Saarakkala[1,2]
[1]University of Oulu, Oulu, Finland [2]Oulu University Hospital, Oulu, Finland [3]Aalto University, Finland

Abstract

This paper addresses the challenge of localization of anatomical landmarks in knee X-ray images at different stages of osteoarthritis (OA). Landmark localization can be viewed as regression problem, where the landmark position is directly predicted by using the region of interest or even full-size images leading to large memory footprint, especially in case of high resolution medical images. In this work, we propose an efficient deep neural networks framework with an hourglass architecture utilizing a soft-argmax layer to directly predict normalized coordinates of the landmark points. We provide an extensive evaluation of different regularization techniques and various loss functions to understand their influence on the localization performance. Furthermore, we introduce the concept of transfer learning from low-budget annotations, and experimentally demonstrate that such approach is improving the accuracy of landmark localization. Compared to the prior methods, we validate our model on two datasets that are independent from the train data and assess the performance of the method for different stages of OA severity. The proposed approach demonstrates better generalization performance compared to the current state-of-the-art.

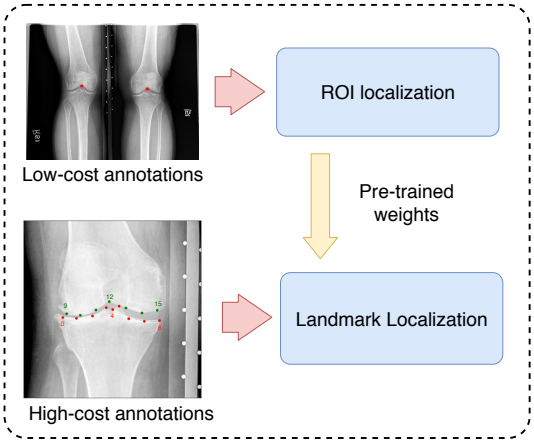


Figure 1. Graphical illustration of our approach. At the first stage, the knee joint area localization model is trained using low-cost annotations. At the second stage, we leverage the weights of the model pre-trained using the low-cost annotations and train a model that localizes 16 individual landmarks. The numbers in the figure indicate the landmark ID (best viewed on screen). The tibial landmarks are displayed in red and numbered from 0 to 8 (left-to-right). Femoral landmarks are displayed in green and numbered from 9 to 15 (left-to-right).

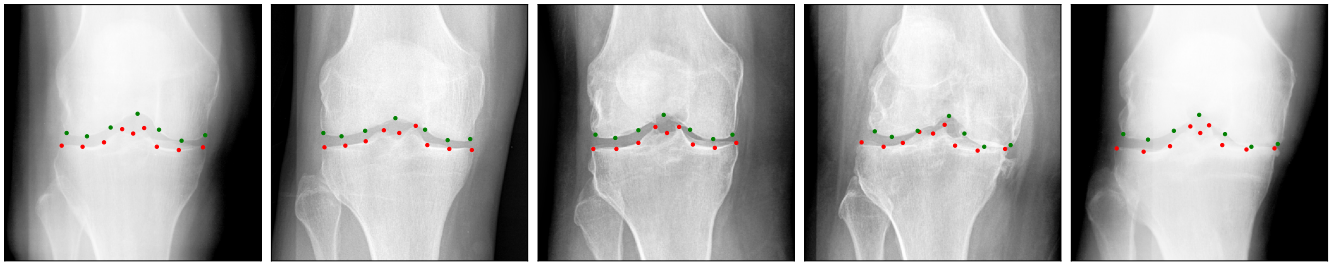
1. Introduction

Anatomical landmark localization is a challenging problem that appears in many medical image analysis problems [31]. One particular realm where the localization of landmarks is of high importance is the analysis of knee plain radiographs at different stages of osteoarthritis (OA) – the most common joint disorder and 11^{th} highest disability factor in the world [2].

In knee OA research field, as well as in the other domains, two sub-tasks that form a typical pipeline for landmark localization can be defined: the region of interest (ROI) localization and the landmark localization itself [41]. In knee radiographs, the former one is typically applied in the analysis of the whole knee images [3, 4, 28, 36, 38], while the latter is used for bone shape and texture analyses [6, 19, 34]. Furthermore, Tiulpin *et al.* also used the landmark localization for image standardization applied after the ROI localization step [36, 37].

Manual annotation of knee landmarks is not a trivial problem without the knowledge of knee anatomy, and it becomes even more challenging when the severity of OA increases. In particular, it makes the annotation process of fine-grained bone edges and tibial spines intractable and time consuming. In Fig. 2, we show the examples of annotations of the landmarks for each stage of OA severity graded according to the gold-standard Kellgren-Lawrence system (grading from 0 to 4) [20]. It can be seen from this figure that when the severity of the disease progresses, bone spurs (osteophytes) and the general bone deformity affect the appearance of the image. Other factors, such as X-ray beam angle are also known to have impact on the image appearance [22].

In this paper, we propose a novel Deep Learning based



(a) KL 0 (b) KL 1 (c) KL 2 (d) KL 3 (e) KL 4

Figure 2. Typical examples of knee joint radiographs at different stages of osteoarthritis severity with overlayed landmarks. Here, the images are cropped to 140×140 mm regions of interest. KL≥ 2 indicates radiographic osteoarthritis. This figure is best viewed on screen.

framework for localization of anatomical landmarks in knee plain radiographs and validate its generalization performance. First, we train a model to localize ROIs in a bilateral radiograph using low-cost labels, and subsequently, train a model on the localized ROIs to predict the location of 16 anatomical landmarks in femur and tibia. Here, we utilize transfer learning and use the model weights from the first step of our pipeline for initialization of the second-stage model. The proposed approach is schematically illustrated in Fig. 1.

Our method is based on the hourglass convolutional network [27] that localizes the landmarks in a weakly-supervised manner and subsequently uses the soft-argmax layer to directly estimate the location of every landmark point. To summarize, the contributions of this study are the following:

- We leverage recent advances in landmark detection using hourglass networks and combine the best design choices in our method.
- For the first time, we propose to use MixUp [42] data augmentation principle for anatomical landmark localization and perform a thorough ablation study for the knee radiographs.
- We demonstrate an effective strategy of enhancing the performance of our landmark localization method by pre-training it on low-budget landmark annotations.
- We evaluate our method on two independent datasets and demonstrate better generalization ability of the proposed approach compared to the current state-of-the-art baseline.
- The pre-trained models, source code and the annotations performed for the Osteoarthritis Initiative (OAI) dataset are publicly available at http://https://github.com/MIPT-Oulu/KNEEL.

2. Related Work

In the literature, there exist only a few studies specifically focused on localization of landmarks in plain knee radiographs. Specifically, the current state-of-the-art was proposed by Lindner *et.al* [24, 25] and it is based on a combination of random forest regression voting (RFRV) with constrained local models (CLM) fitting.

There are several methods focusing solely on the ROI localization. Tiulpin *et al.* [39] proposed a novel anatomical proposal method to localize the knee joint area. Antony *et al.* [3] used fully convolutional networks for the same problem. Recently, Chen *et al.* [9] proposed to use object detection methods to measure the knee OA severity.

The proposed approach is related to the regression-based methods for keypoint localization [41]. We utilize an hourglass network which is an encoder-decoder model initially introduced for human pose estimation [27] and address both ROI and landmark localization tasks. Several other studies in medical imaging domain also leveraged a similar approach by applying U-Net [33] to the landmark localization problem [12, 31]. However, the encoder-decoder networks are computationally heavy during the training phase since they regress a tensor of high-resolution heatmaps which is challenging for medical images that are typically of a large size. It is notable that decreasing the image resolution could negatively impact the accuracy of landmark localization. In addition, most of the existing approaches use a refinement step which makes the computational burden even harder to cope with. Nevertheless, hourglass CNNs are widely used in human pose estimation [27] due to a possibility of lowering down the resolution and the absence of precise ground truth.

More similar to our approach, Honari *et al.* [18] recently leveraged deep learning and applied soft-argmax layer to the feature maps of the full image resolution to improve landmark localization performance leading to remarkable results. However, such strategy is computationally heavy for medical images due to their high resolution. In contrast, we first moderately reduce the image resolution by embed-

ding it into a feature space, utilize an hourglass module to process the obtained feature maps at all scales, and eventually apply the soft-argmax operator that makes the proposed configuration more applicable to high-resolution images allowing to get sub-pixel accurate landmark coordinates.

3. Method

3.1. Network architecture

Overview. Our model comprises several architectural components of modern hourglass-like encoder-decoder models for landmark localization. In particular, we utilize the hierarchical multi-scale parallel (HMP) residual block [7] which improves the gradient flow compared to the traditional bottleneck layer described in: [17, 27]. The HMP block structure is illustrated in Fig. 3.

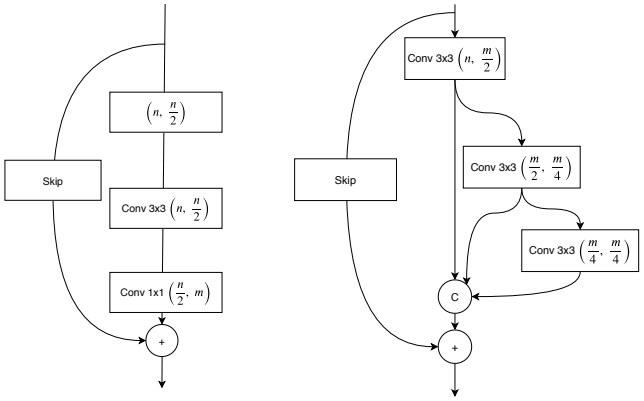


Figure 3. Graphical illustration of the difference between the bottleneck residual block [27, 17] (a) and the HMP residual block [7] (b). Here, n and m indicate the number of input and output feature maps, respectively. Skip connection representing 1×1 convolution is applied if $n \neq m$.

The architecture of the proposed model is represented in Fig. 4. In general, our model comprises three main components: entry block, hourglass block, and output block. The whole network is parameterized by two hyperparameters – width N and depth d, where the latter is related to the number of max-pooling steps in the hourglass block. In our experiments we found the width of $N = 24$ and the depth of $d = 6$ to be optimal to maintain both high accuracy and speed of computations.

Entry block. Similar to the original hourglass model [27] we apply a 7×7 convolution with stride 2 and zero padding of 3 and pass the results into a residual module. Further, we use a 2×2 max-pooling and utilize three residual modules before the hourglass block. This block allows to simultaneously downscale the image 4 times and obtain representative feature embeddings suitable for multi-scale processing performed in the hourglass block.

Hourglass block. This block starts with a 2×2 max-pooling and recursively repeats dual-path structure d times as can be seen in Fig. 4. In particular, each level of the hourglass block starts with a 2×2 max-pooling subsequently followed by 3 HMP residual blocks. At the next stage, the representations from the current level i are passed to the next hourglass' level $i + 1$ and also passed forward to be summed with the up-sampled outputs of the hourglass level $i + 1$. Since spatial resolution of the feature maps at level i and $i + 1$ is different, the nearest-neighbours up-sampling is used [27]. At level d, we simply feed the representations into the HMP block instead of the next hourglass level due to the reached limit of hourglass' depth.

Output block. The final block of the model uses the representations coming from the hourglass module and sequentially applies two blocks of dropout ($p = 0.25$) and 1×1 convolutional block with batch normalization and ReLU. At the final stage, a 1×1 convolution and soft-argmax [8] are utilized to regress the coordinates of each landmark point.

Soft-argmax. Since soft-argmax is an important component of our model, we review its formulation in this paragraph. This operator can be defined as a sequence of two steps, where the first one calculates the spatial softmax for pixel (i, j):

$$\Phi(\beta, \mathbf{h}, i, j) = \frac{\exp[\beta \mathbf{h}_{ij}]}{\sum_{k=0}^{W-1} \sum_{l=0}^{H-1} \exp[\beta \mathbf{h}_{kl}]} \tag{1}$$

At the next stage, the obtained spatial softmax is multiplied by the expected value of landmark coordinate at every pixel:

$$\Psi_d(\mathbf{h}) = \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} \mathbf{W}_{ij}^{(d)} \Phi(\beta, \mathbf{h}, i, j), \tag{2}$$

where

$$\mathbf{W}_{ij}^{(x)} = \frac{i}{W}, \mathbf{W}_{ij}^{(y)} = \frac{j}{H}. \tag{3}$$

3.2. Loss function

We assessed various loss functions for training our model and finalized our choice at wing loss [15] that is closely related to L_1 loss. However, in the case of wing loss, the errors in a small vicinity of 0 – $(-w, w)$ are better amplified due to the logarithmic nature of the function:

$$L(y, \hat{y}) = \begin{cases} w \log\left(1 + \frac{1}{\epsilon}|y - \hat{y}|\right) & |y - \hat{y}| < w \\ |y - \hat{y}| - C & \text{otherwise} \end{cases}, \tag{4}$$

where y – is a ground truth, $\hat{y}$ – prediction, $(-w, w)$ – range of non-linear part of the loss, C – constant smoothly linking the linear and non-linear parts.

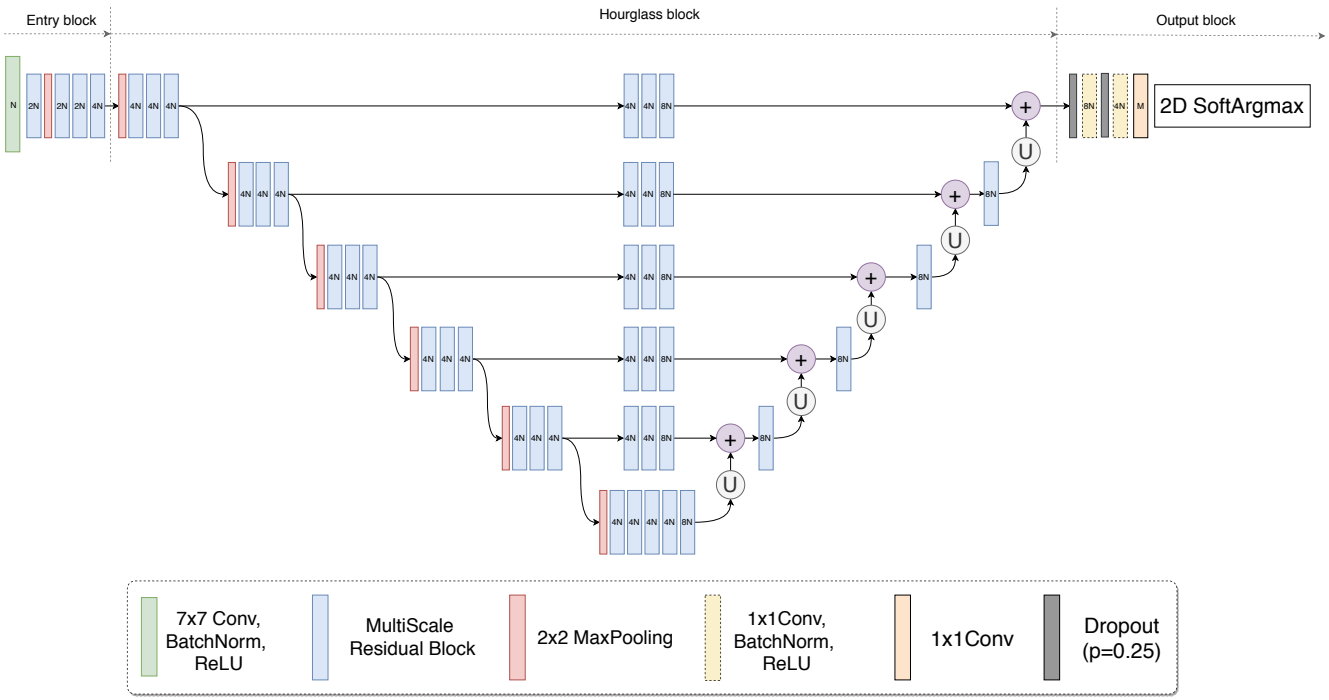


Figure 4. Model architecture with an hourglass block of depth $d = 6$. Here, N is a width of the network and M is the number of output landmarks.

3.3. Training techniques

MixUp We use a MixUp technique [42] to improve the performance of our method. In particular, MixUp mixes the data inputs x_1 and x_2, the corresponding keypoint arrays p_1 and p_2:

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha) \tag{5}$$
$$\lambda' = \max(\lambda, 1 - \lambda) \tag{6}$$
$$x' = \lambda' x_1 + (1 - \lambda') x_2 \tag{7}$$
$$p' = \lambda' p_1 + (1 - \lambda') p_2, \tag{8}$$

thereby augmenting the dataset with the new interpolated examples. Our implementation of mixup does not differ from the one proposed in the original work[1] and we do not compute the mixed targets p'. In contrast, we rather optimize the following loss function calculated mini-batch-wise:

$$L'(x_1, x', p_1, p_2) = \lambda L(p_1, o_1) + (1 - \lambda) L(p_2, o'), \tag{9}$$

where o_1 and o' are the outputs of the network for x_1 and x', respectively. Here, the points p_2 for every point p_1 are generated by a simple mini-batch shuffling.

Data Augmentation. Medical images can vary in appearance due to different data acquisition settings or patient-related anatomical features. To tackle the issue of limited data, we applied the data augmentation. We use geometric and textural augmentations similarly to to the face landmark detection problem [16]. The former included all classes of homographic transformations while the latter included gamma correction, salt and pepper, blur (both median and gaussian) and the addition of a gaussian noise. Interestingly, the homographic transformations were shown effective in improving, for example, self-supervised learning [23, 26], however only more narrow class of transformation (affine) has been applied to the landmark localization [16] in faces.

Transfer learning from low-budget annotations. As shown in Fig. 1, the problem of localizing the landmarks comprises two stages: identification of the ROI and the actual landmark localization. We previously mentioned the two classes of labels that are needed to train such a pipeline: low-cost ($1 - 2$ points / image) and high-cost labels ($2+$ points). The low-cost labels can be noisy / inaccurate and are quick to produce, while the high-cost labels require the expert knowledge. In this work, we first train the ROI localization model (1 landmark per leg) on the low-cost labels – knee joint centers (see Fig. 1) and then re-use the pre-trained weights from this stage to train the landmark localization model (16 landmarks per knee joint).

[1] https://github.com/facebookresearch/mixup-cifar10

4. Experiments

4.1. Datasets

Annotation Process For all the following datasets, we applied the same annotations process. Firstly, for all the images in all the datasets we run BoneFinder tool (see Sec. 4.2). At the second stage, for every image, a person experienced in knee anatomy and OA manually refine all the landmark points. In Fig. 1, we highlight the numbering of the landmarks that we use in this paper. Specifically, we marked the corner landmarks in tibia from 0 to 8 and in femur from 9 to 15 (lateral to medial). To perform the annotations, we used VGG image annotation tool [14].

OAI. We trained our model and performed model selection using the images from Osteoarthritis Initiative (OAI) dataset[2]. Roughly 150 knee joint images per KL grade were sampled to be included into the dataset. The final dataset size comprised 748 knee joints in total. In the case of the ROI localization, we used a half of the image that corresponded to each knee.

Dataset A. These data were collected at our hospital (Oulu University Hospital, Finland) [32], and thus, it comes from a completely different population than OAI (from USA). It includes the images from 81 subjects, and KL grade-wise the data have the following distribution: 4 knees with KL 0, 54 knees with KL 1, 49 knees with KL 2, 29 knees with KL 3 and 25 knees with KL 4. From this dataset, we excluded 1 knee due to an implant, thereby using 161 knees for testing of our model.

Dataset B. This dataset was also acquired from our hospital (Oulu University Hospital, Finland; ClinicalTrials.gov ID: NCT02937064) and included originally 107 subjects. Out of these, 5 knee joints were excluded, thereby making a dataset of 209 knees (4 implants and 1 due to error during the annotation process). With respect to OA severity, these data had 35 cases with KL 0, 84 with KL 1, 51 with KL 2, 37 with KL 3 and 2 with KL 4. This dataset was also used solely for testing of our model.

4.2. Baseline methods

We used several baseline methods at the model selection phase and one strong pre-trained baseline method at the test phase. In particular, we used Active Appearance Models [10] and Constrained Local Models [11] with both Image Gradient Orientations (IGO) [40] and Local Binary Patterns Features (LBP) [29]. Our implementation is based on the available methods with default hyperparameters from the Menpo library [1].

At the test phase, we used pre-trained RFRV-CLM method [25] implemented in BoneFinder tool. Here, the RFRV-CLM model was trained on 500 images from OAI dataset. However we did not have access to the train data to assess which samples were used for training this method, therefore, we used this tool only for testing on datasets A and B.

[2] https://oai.epi-ucsf.org/datarelease/

4.3. Implementation Details

Ablation experiments All our ablation experiments were conducted on the same 5-fold patient-wise cross-validation split stratified by a KL grade to ensure equal distribution of different stages of OA severity. Both ROI and landmark localization models were trained using the same split.

During the training, we used exactly the same hyperparameters for all the experiments. In particular, we used $N = 24$ and $d = 6$ for our network. The learning rate and the batch size were fixed to $1e-3$ and 16, respectively. In some of our experiments where the weight decay was used, we set it to $1e-4$. All the models were trained with Adam optimizer [21]. The pixel spacing for ROI localization was set to 1 mm and for the landmark localization to 0.3 mm. We used bi-linear interpolation for image resizing.

All the ablation experiments were conducted solely on landmark localization task and eventually, after selecting the best configuration, we used it for training the ROI localization model due to the similarity of the tasks. We used the ground truth annotations to crop the 140×140 mm ROIs around the tibial center (landmark 4 in Fig. 1) to create the data for model selection and training the landmark localization model. In our experiments, we flipped all the left ROI images to look like the right ones, however this strategy was not applied for the ROI localization task.

When performing the fine-tuning of landmark localization model using the pre-trained weights of the ROI localization model, we simply initialized all the layers of the former with the weights of the latter one. We note here that the last layer was initialized randomly and we did not freeze the pre-trained part for simplicity.

In our experiments, we used PyTorch v1.1.0 [30] on a single Nvidia GTX1080Ti. For data augmentation, we used SOLT library [35]. For training AAM and CLM, we used Menpo [1], as mentioned earlier.

Evaluation and Metrics To assess the results of our method, we used multiple metrics and evaluation strategies. Firstly, we performed the ablation experiments and used the landmarks $0, 8, 9, 15$ for evaluation of the results (see Fig. 1). At the test time, when comparing the performance of the full system, we used an extended set of landmarks for evaluation – $0, 4, 8, 9, 12, 15$. The intuition here is to compare the landmark methods on those landmark points that are the most crucial in applications (tibial corners for landmark localization as well as tibial and femoral centers for the ROI localization). Besides, we excluded all the knees with implants from the evaluation.

As as the main metric for comparison, we used Percentage of Correct Keypoints (PCK) @ r to compare the landmark localization methods. This metric shows the percentage of points that fall within the neighborhood of a ground truth landmark having the radius r (recall at different precision thresholds). In our experiments, we used r of 1 mm, 1.5 mm, 2 mm and 2.5 mm for quantitative comparison.

Finally, we also assessed the amount of outliers in the landmark localization task. An outlier was defined as a landmark that do not fall within the 10 mm neighbourhood of the ground truth landmark. This value was computed for all the landmark points in contrast to PCK.

4.4. Ablation Study

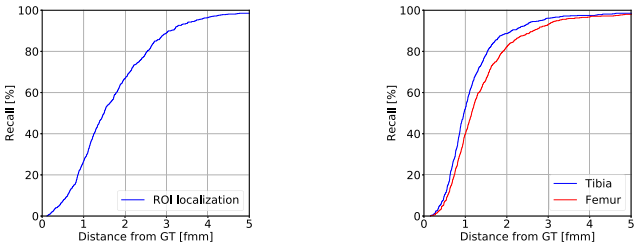
Conventional approaches. We first investigated the conventional approaches for landmark localization. The benchmarks of AAM and CLM with IGO and LBP features with default hyperparameters from Menpo [1] showed satisfactory results. The best model here was CLM with IGO features (Tab. 1).

Loss Function. In the initial experiments with our model we assessed different loss functions (see Tab. 1). In particular, we used L_2,L_1, wing [15] and elastic loss (sum of L_2 and L_1 losses). Besides, we also utilized a recently introduced general adaptive robust loss with the default hyperparameters [5]. Our experiments showed that wing loss with the default hyperparameters as in the original paper ($w = 15$ and $C = 3$), produces the best results.

Effect of Multi-scale Residual Blocks. The experiments done for loss functions were conducted using the HMP block. However, it is worth to assess the added value of this block compare to the bottleneck residual block. Tab. 1 demonstrates that the bottleneck residual block ("Wing + regular res. block" of the Table) fell behind of HMP ("Wing loss") in terms of PCK.

MixUp vs. Weight Decay After observing that the wing loss and HMP block yield the best default configuration, we experimented with various forms of regularization. In this series of experiments, we used our default configuration and applied MixUp with different α. Our experiments showed that using MixUp the default configuration and weight decay degrades the performance (Tab. 1). However, MixUp itself is also a powerful regularizer, therefore, we conducted the experiments without weight decay (marked as *no wd* in Tab. 1). Interestingly, setting weight decay to 0 increases the performance of our model with any α. To assess the strength of regularization, we also conducted an experiment with $\alpha = 0.75$ (best) and without dropout. We observed that having dropout helps MixUp.

CutOut vs. Target Jitter Besides MixUp, we tested two other data augmentation techniques – cutout [13] and noise addition to the ground truth annotations during the training (uniform distribution, ± 1 pixel). We observed that the latter did not improve the results of our configuration with MixUp, however the former helped to lower down the amount of outliers twice while yielding nearly the same localization performance. This configuration had a cutout of 10% of the image. These results are also presented in Tab. 1.



(a) (b)

Figure 5. Cumulative plots reflecting the performance of ROI (a) and landmark (b) localization methods on cross-validation. ROI localization was assessed at the pixel spacing of 1 mm and the landmark localization at 0.3 mm, respectively. GT indicates ground truth.

Transfer Learning from Low-cost Labels. At the final stage of our experiments, we used the best configuration that included the wing loss, MixUp with $\alpha = 0.75$, weight decay of 0 and 10% cutout to train the ROI localization model. Essentially, both of these methods are landmark localization approaches, therefore, in our cross-validation experiments, we also assessed the performance of ROI localization using PCK. In our experiments, we found that pre-training of the landmark localization model on the ROI localization task significantly increases the performance of the former (see the last row of Tab. 1). The performance of both these models on cross-validation is presented in Fig. 5. Quantitatively, ROI localization model yielded PCK of 26.60%, 50.27%, 66.71%, 79.14% at 1 mm, 1.5 mm, 2 mm and 2.5 mm thresholds, respectively and had 0.13% outliers.

4.5. Test datasets

Testing on the full datasets Testing of our model was conducted on datasets A and B, respectively. We provide the quantitative results in Tab. 2. In this table, we present two versions of our pipeline, one is a single stage, where the landmark localization follows directly after the ROI localization step, and also a two-stage pipeline that includes ROI localization as a first step, initial inference of the landmark points as a second step, and re-centering of the ROI to the predicted tibial center and a second pass of landmark localization model as a third step.

Setting	1 mm	1.5 mm	2 mm	2.5 mm	% out
AAM (IGO [40])	7.29 ± 4.06	17.18 ± 5.39	28.07 ± 5.29	39.51 ± 6.33	7.49
AAM (LBP [29])	2.41 ± 0.19	8.02 ± 1.13	15.17 ± 3.12	24.33 ± 4.73	9.22
CLM (IGO [40])	24.53 ± 3.31	39.84 ± 4.92	50.60 ± 3.69	61.43 ± 4.25	3.61
CLM (LBP [29])	2.67 ± 1.51	10.03 ± 3.21	18.65 ± 5.77	28.81 ± 5.58	9.36
L2 loss	0.00 ± 0.00	0.00 ± 0.00	0.07 ± 0.09	0.07 ± 0.09	92.78
L1 loss	17.45 ± 5.20	45.45 ± 5.48	66.11 ± 5.39	80.08 ± 3.78	2.67
Robust loss [5]	13.97 ± 0.47	35.83 ± 1.70	57.35 ± 1.89	72.06 ± 1.89	4.68
Elastic loss	4.14 ± 3.40	13.97 ± 7.66	27.21 ± 9.74	41.58 ± 10.59	9.36
Wing loss [15]	31.68 ± 5.10	61.83 ± 7.09	78.68 ± 5.58	87.50 ± 3.31	2.14
Wing + regular res. block	25.74 ± 3.31	55.48 ± 3.97	73.46 ± 3.69	83.82 ± 3.03	2.67
Wing + mixup $\alpha = 0.1$	27.54 ± 0.19	58.42 ± 1.70	77.21 ± 1.42	87.17 ± 0.57	2.27
Wing + mixup $\alpha = 0.2$	29.88 ± 4.25	58.96 ± 2.84	78.07 ± 6.05	86.16 ± 3.50	2.94
Wing + mixup $\alpha = 0.5$	29.61 ± 1.42	59.36 ± 3.03	77.81 ± 3.78	86.30 ± 2.55	2.67
Wing + mixip $\alpha = 0.75$	30.75 ± 3.40	59.63 ± 4.92	77.07 ± 5.20	86.36 ± 2.84	3.48
Wing + mixup $\alpha = 0.1$ (no wd)	34.89 ± 5.29	63.64 ± 7.56	81.15 ± 5.48	89.24 ± 3.12	1.47
Wing + mixup $\alpha = 0.2$ (no wd)	35.16 ± 5.86	64.17 ± 7.00	82.15 ± 5.58	89.91 ± 4.25	1.34
Wing + mixup $\alpha = 0.5$ (no wd)	36.30 ± 6.33	65.04 ± 6.33	81.82 ± 4.16	89.91 ± 2.55	1.47
Wing + mixup $\alpha = 0.75$ (no wd)	37.97 ± 5.48	67.45 ± 4.25	82.02 ± 1.80	90.51 ± 0.95	1.60
Wing + mixup $\alpha = 0.75$ (no wd, no dropout)	37.10 ± 5.39	65.64 ± 3.97	81.75 ± 4.44	89.30 ± 3.21	1.47
Wing + mixup $\alpha = 0.75$ + jitter (no wd)	36.63 ± 4.16	65.98 ± 5.58	83.09 ± 3.88	90.84 ± 3.31	1.60
Wing + mixup $\alpha = 0.75$ + cutout 5% (no wd)	34.96 ± 3.69	63.30 ± 6.14	80.15 ± 4.06	89.30 ± 1.32	1.07
Wing + mixup $\alpha = 0.75$ + cutout 10% (no wd)	37.83 ± 4.35	65.78 ± 4.35	81.35 ± 3.50	90.24 ± 1.51	**0.53**
Wing + mixup $\alpha = 0.75$ + cutout 25% (no wd)	35.56 ± 3.97	62.50 ± 5.01	80.01 ± 4.06	88.50 ± 2.84	0.94
Wing + mixup $\alpha = 0.75$ + cutout 10% (no wd, finetune)	$\mathbf{45.92 \pm 8.79}$	$\mathbf{72.39 \pm 8.60}$	$\mathbf{85.36 \pm 4.63}$	$\mathbf{90.91 \pm 3.21}$	1.34

Table 1. Results of the model selection for high-cost annotations on the OAI dataset. The values of PCK/recall (%) at different precision are shown as average and standard deviation for the landmarks 0, 8, 9, 15, while the amount of outliers is calculated for all the landmarks. The comparison is done at 0.3 mm image resolution (pixel spacing). Best results are highlighted in bold.

Dataset	Method	Precision				% out
		1 mm	1.5 mm	2 mm	2.5 mm	
A	BoneFinder [25]	$\mathbf{48.45 \pm 2.64}$	$\mathbf{59.63 \pm 3.51}$	78.26 ± 7.03	89.13 ± 3.95	**0.00**
	Ours 1-stage	12.73 ± 2.20	46.89 ± 5.71	78.57 ± 1.32	90.99 ± 1.32	1.24
	Ours 2-stage	14.60 ± 4.83	47.52 ± 2.20	$\mathbf{78.88 \pm 0.88}$	$\mathbf{93.48 \pm 0.44}$	0.62
B	BoneFinder [25]	2.87 ± 3.38	13.64 ± 10.49	43.78 ± 21.31	68.90 ± 20.98	**0.00**
	Ours 1-stage	9.33 ± 1.01	42.58 ± 1.35	74.40 ± 1.69	91.63 ± 1.69	0.48
	Ours 2-stage	$\mathbf{11.24 \pm 0.34}$	$\mathbf{44.98 \pm 0.68}$	$\mathbf{75.12 \pm 2.71}$	$\mathbf{92.11 \pm 0.34}$	0.48

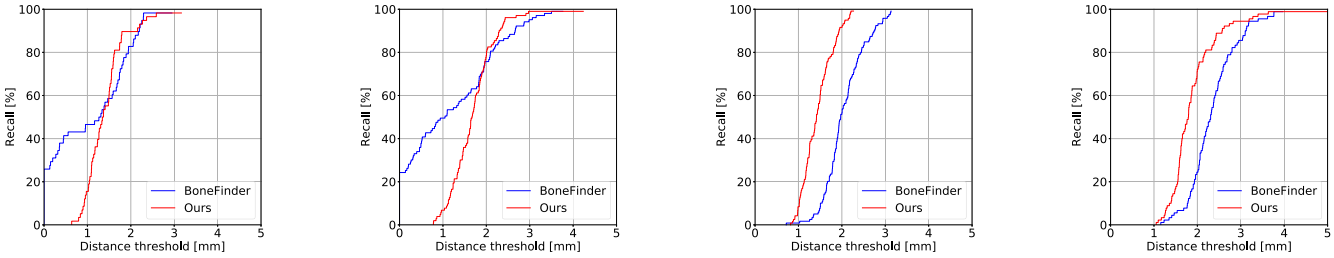
Table 2. Test set results and comparison to the state-of-the-art method (RFRV-CLM-based BoneFinder tool) by Lindner *et al*. [25]. Reported percentage of outliers is calculated for *all* landmarks, while the PCK/recall values (%) are calculated as the average for the landmarks 0, 4, 8, 9, 12, and 15. Best results per dataset are highlighted in bold. It should be noted that BoneFinder operated with the full image resolution while our method performed ROI localization at 1 mm and landmark localization at 0.3 mm resolutions, respectively.

Testing with Respect to the presence of Radiographic Osteoarthritis To better understand the behaviour of our model on the test datasets, we investigated the performance of our 2-stage pipeline and BoneFinder for cases having KL < 2 and KL ≥ 2, respectively. These results are presented in Fig. 6. Our method performs on par with BoneFinder for Dataset A and even exceeds its localization performance for precision thresholds above 2 mm for radiograhic OA. In Dataset B, on average, our method performs better than BoneFinder when both methods are benchmarked for both non-OA and OA cases. To provide better insights into the performance of our method for different stages of OA severity, we show examples of landmark localization done by our method, BoneFinder and manually (Fig. 7).

5. Conclusions

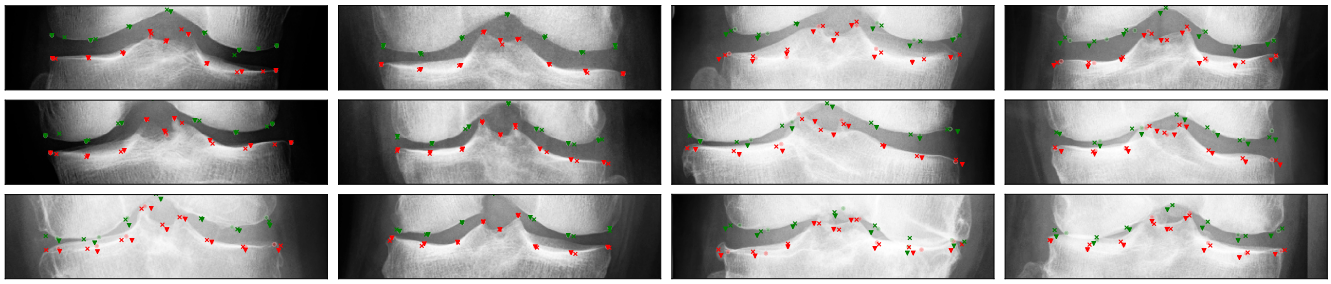
In this paper, we addressed the problem of anatomical landmark localization in knee radiographs. We proposed a new method that leverages the power of latest advances in landmark localization and pose estimation.

Compared to the current state-of-the-art [24, 25], our method generalized better to the unseen test datasets that had completely different acquisition settings and patient populations. Consequently, these results suggest that our



(a) Dataset A (no OA) (b) Dataset A (OA) (c) Dataset B (no OA) (d) Dataset B (OA)

Figure 6. Cumulative distribution plots of localization errors for our two-stage method and BoneFinder [25, 24] for cases with and without radiographic OA on datasets A and B, respectively.



(a) Dataset A (worst) (b) Dataset A (best) (c) Dataset B (worst) (d) Dataset B (best)

Figure 7. Examples of predictions on datasets A and B (worst and best cases). We visualized ground truth landmarks as circles. Predictions made by our method are shown using crosses and predictions made by BoneFinder are shown using triangles. Red and green show the landmarks for tibia and femur, respectively. Best and worst cases were selected based on the average total error of *our method* per group. The width of every example is 115 mm. The first row contains examples having KL 0 or 1, the second row contains examples with KL 2 and the third row with KL 3.

approach may be easily applicable to various tasks in clinical and research settings.

Our study has still some limitations. Firstly, the comparison with BoneFinder should ideally be conducted when it is trained on the same 0.3 mm resolution data with the same KL grade-wise stratification, or at full image resolution. However, we did not have access to the training code of BoneFinder, thereby, leaving more systematic comparison to future studies. Another limitation of this study is the ground truth annotation process. Specifically, we used BoneFinder to pre-annotate the landmarks for all the images in both train and test sets. In theory, this might give an advantage to BoneFinder compared to our method. On the other hand, all the landmarks were still manually refined, which should decrease this advantage.

The core methodological novelties of the study were in adapting the MixUp, soft-argmax layer and transfer learning from low-cost annotations for training our model. We think that the latter has applications in other, even non-medical domains, such as human pose estimation and facial landmark localization. It was shown that compared to RFRV-CLM, Deep Learning methods scale with the amount of training data, and therefore, we also expect our method to yield even better results when it is trained on a larger datasets [12]. Besides, we also expect semi-supervised learning [18] to help in this task.

6. Acknowledgements

This study was supported by KAUTE foundation, Infotech Oulu, University of Oulu strategic funding and Sigrid Juselius Foundation.

The OAI is a public-private partnership comprised of five contracts (N01- AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health.

Development and maintenance of VGG Image Annotator (VIA) is supported by EPSRC programme grant Seebibyte: Visual Search for the Era of Big Data (EP/M013774/1).

We thank Dr. Claudia Lindner for providing BoneFinder.

References

[1] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 679–682. ACM, 2014. 5, 6

[2] K. D. Allen and Y. M. Golightly. Epidemiology of osteoarthritis: state of the evidence. *Current opinion in rheumatology*, 27(3):276, 2015. 1

[3] J. Antony, K. McGuinness, K. Moran, and N. E. OConnor. Automatic detection of knee joints and quantification of knee osteoarthritis severity using convolutional neural networks. In *International conference on machine learning and data mining in pattern recognition*, pages 376–390. Springer, 2017. 1, 2

[4] J. Antony, K. McGuinness, N. E. O'Connor, and K. Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1195–1200. IEEE, 2016. 1

[5] J. T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4331–4339, 2019. 6, 7

[6] A. Brahim, R. Jennane, R. Riad, T. Janvier, L. Khedher, H. Toumi, and E. Lespessailles. A decision support tool for early detection of knee osteoarthritis using x-ray imaging and machine learning: Data from the osteoarthritis initiative. *Computerized Medical Imaging and Graphics*, 73:11–18, 2019. 1

[7] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3706–3714, 2017. 3

[8] O. Chapelle and M. Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information retrieval*, 13(3):216–235, 2010. 3

[9] P. Chen, L. Gao, X. Shi, K. Allen, and L. Yang. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 2019. 2

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 23(6):681–685, 2001. 5

[11] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *Bmvc*, page 3. Citeseer, 2006. 5

[12] A. K. Davison, C. Lindner, D. C. Perry, W. Luo, T. F. Cootes, et al. Landmark localisation in radiographs using weighted heatmap displacement voting. In *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*, pages 73–85. Springer, 2018. 2, 8

[13] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[14] A. Dutta and A. Zisserman. The VIA annotation software for images, audio and video. *arXiv preprint arXiv:1904.10699*, 2019. 5

[15] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018. 3, 6, 7

[16] Z.-H. Feng, J. Kittler, and X.-J. Wu. Mining hard augmented samples for robust facial landmark localization with cnns. *IEEE Signal Processing Letters*, 26(3):450–454, 2019. 4

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[18] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, and J. Kautz. Improving landmark localization with semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2018. 2, 8

[19] T. Janvier, H. Toumi, K. Harrar, E. Lespessailles, and R. Jennane. Roi impact on the characterization of knee osteoarthritis using fractal analysis. In *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 304–308. IEEE, 2015. 1

[20] J. Kellgren and J. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494, 1957. 1

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] M. Kothari, A. Guermazi, G. von Ingersleben, Y. Miaux, M. Sieffert, J. E. Block, R. Stevens, and C. G. Peterfy. Fixed-flexion radiography of the knee provides reproducible joint space width measurements in osteoarthritis. *European radiology*, 14(9):1568–1573, 2004. 1

[23] Z. Laskar, I. Melekhov, H. R. Tavakoli, J. Ylioinas, and J. Kannala. Geometric image correspondence verification by dense pixel matching. *arXiv preprint arXiv:1904.06882*, 2019. 4

[24] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1862–1874, 2014. 2, 7, 8

[25] C. Lindner, S. Thiagarajah, J. M. Wilkinson, G. A. Wallis, T. F. Cootes, arcOGEN Consortium, et al. Accurate bone segmentation in 2d radiographs using fully automatic shape model matching based on regression-voting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 181–189. Springer, 2013. 2, 5, 7, 8

[26] I. Melekhov, A. Tiulpin, T. Sattler, M. Pollefeys, E. Rahtu, and J. Kannala. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1034–1042. IEEE, 2019. 4

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2, 3

[28] B. Norman, V. Pedoia, A. Noworolski, T. M. Link, and S. Majumdar. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *Journal of digital imaging*, 32(3):471–477, 2019. 1

[29] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(7):971–987, 2002. 5, 7

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS Workshop Autodiff*, December 2017. 5

[31] C. Payer, D. Štern, H. Bischof, and M. Urschler. Integrating spatial configuration into heatmap regression based cnns for landmark localization. *Medical Image Analysis*, 54:207–219, 2019. 1, 2

[32] J. Podlipská, A. Guermazi, P. Lehenkari, J. Niinimäki, F. W. Roemer, J. P. Arokoski, P. Kaukinen, E. Liukkonen, E. Lammentausta, M. T. Nieminen, et al. Comparison of diagnostic performance of semi-quantitative knee ultrasound and knee radiography with mri: Oulu knee osteoarthritis study. *Scientific reports*, 6:22365, 2016. 5

[33] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[34] J. Thomson, T. ONeill, D. Felson, and T. Cootes. Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 127–134. Springer, 2015. 1

[35] A. Tiulpin. Solt: Streaming over lightweight transformations. https://github.com/MIPT-Oulu/solt, 2019. 5

[36] A. Tiulpin, S. Klein, S. Bierma-Zeinstra, J. Thevenot, E. Rahtu, J. van Meurs, E. H. Oei, and S. Saarakkala. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *arXiv preprint arXiv:1904.06236*, 2019. 1

[37] A. Tiulpin and S. Saarakkala. Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. *arXiv preprint arXiv:1907.08020*, 2019. 1

[38] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific reports*, 8(1):1727, 2018. 1

[39] A. Tiulpin, J. Thevenot, E. Rahtu, and S. Saarakkala. A novel method for automatic localization of joint area on knee plain radiographs. In *Scandinavian Conference on Image Analysis*, pages 290–301. Springer, 2017. 2

[40] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Subspace learning from image gradient orientations. *IEEE transactions on pattern analysis and machine intelligence*, 34(12):2454–2466, 2012. 5, 7

[41] Y. Wu and Q. Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2):115–142, 2019. 1, 2

[42] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 4