DATA-EFFICIENT KNEE ANATOMICAL LANDMARK LOCALIZATION USING DEEP LEARNING

Raimo Niemelä
Master's thesis

Biomedical Engineering
Faculty of Medicine
University of Oulu
2021

Abstract

Niemelä Raimo (2021), Data-efficient Knee Anatomical Landmark Localization using Deep Learning. Faculty of Medicine, University of Oulu, Master's Thesis, 49 pages.

Knee osteoarthritis (OA) is the most common musculoskeletal degenerative disease affecting the joints. OA is examined at a doctor's visit and an X-ray image is often used to confirm the diagnosis. There is no treatment available for OA, therefore it is important to diagnose knee osteoarthritis at the earliest possible stage to prevent possible complications.

Traditional methods used by a practitioners do not detect osteoarthritis as early as possible, therefore other methods are needed for early diagnosis. One possibility is to use novel quantitative imaging biomarkers, computation of which often requires precise understanding of the knee anatomy by a computer. More specifically, it is important to locate different areas of the knee according to anatomical atlases and place relevant regions of interest to compute the imaging biomarkers. A state-of-the-art approach for this problem is based on anatomical landmark localization.

In this work, the localization of anatomical landmarks from knee X-rays using deep learning is investigated. To date, statistical methods have been used to localize landmarks, but this work focuses on identification based on deep learning and investigates how the amount of available training data affects performance. The method investigated in the present thesis is based on the KNEEL method developed earlier at the University of Oulu. The aim of this work was to improve this method by adjusting the training parameters and leveraging equivalent regularization for semi-supervised learning. Images from the Osteoarthritis Initiative database were used as material for training and validation.

During the work, it was found that by adjusting the parameters used for training, anatomical landmarks can be localized more accurately than in the original KNEEL method. By adding the equivalent regularization, the accuracy of the localization was increased substantially, and a further enhancement in performance can be observed by utilizing unlabeled data in a semi-supervised learning manner.

The results, developed in this thesis can layer be leveraged in OA research or even clinical practice, where the computation of quantitative imaging biomarkers is important. To our knowledge, this is the first work in OA where SSL and equivariant regularization were used.

Keywords: osteoarthritis, deep learning, knee, supervised learning, semi-supervised learning

Tiivistelmä

Niemelä Raimo (2021), Datatehokas polven anatomisten maamerkkien paikantaminen käyttäen syväoppimista, Lääketieteellinen tiedekunta, Oulun yliopisto, Pro gradu -tutkielma, 49 sivua.

Polven nivelrikko on yleisin niveliin vaikuttava tuki- ja liikuntaelimistöä rappeuttava sairaus. Nivelrikko tutkitaan lääkärikäynnin yhteydessä ja diagnoosi vahvistetaan usein röntgenkuvantamisen avulla. Nivelrikkoon ei ole saatavilla hoitoa, joten on tärkeää diagnosoida polven nivelrikko mahdollisimman varhaisessa vaiheessa mahdollisten komplikaatioiden välttämiseksi.

Perinteiset lääkäreiden käyttämät menetelmät eivät tunnista nivelrikkoa riittävän aikaisin, siksi tarvitaan muita menetelmiä varhaisempaan diagnostiikkaan. Yksi mahdollisuus on käyttää kvantitatiivisia kuvantamisbiomarkkereita, mutta näiden laskemiseksi tietokoneen täytyy ymmärtää anatomisia rakenteita tarkasti. Tarkemmin sanottuna on tärkeää paikantaa polven eri rakenteet ihmisen anatomiasta ja merkitä kiinnostavat rakenteet, jotta kuvantamisbiomarkkerit voidaan laskea. Nykyisin tätä ongelmaa lähestytään anatomisten maamerkkien paikantamisen avulla.

Tässä työssä tutkittiin anatomisten maamerkkien paikantamista polven röntgenkuvista syväoppimisen avulla. Perinteisesti tähän on käytetty staattisia menetelmiä, mutta tässä työssä keskityttiin paikantamiseen käyttäen syväoppimista ja tutkittiin kuinka käytettävissä oleva opetusdatan määrä vaikuttaa suorituskykyyn. Työssä käytetty metodi perustuu aikaisemmin Oulun yliopistossa kehitettyyn KNEEL metodiin. Tämän työn tarkoituksena oli parantaa tätä metodia säätämällä opetusparametreja sekä hyödyntää ekvivalenttia regularisaatiota syväoppimisen yhteydessa. Kuvia The Osteoarthritis Initiative -tietokannasta käytettiin opetukseen ja validointiin.

Työn aikana havaittiin, että säätämällä opetukseen käytettäviä parametrejä, voidaan anatomiset maamerkit paikantaa tarkemmin kuin alkuperäisellä KNEEL metodilla. Ekvivalentin regularisaation lisäämisellä paikantamisen tarkkuus lisääntyi huomattavasti. Suorituskyky parani entisestään käyttämällä annotoimatonta dataa puoli-ohjatun oppimisen yhteydessä.

Tämän opinnäytetyön yhteydessä kehitettyä metodia voidaan käyttää nivelrikon tutkimuksen yhteydessä tai kliinisessä käytössä, missä kvantitatiivisten kuvantamisbiomarkkereiden käyttö on tärkeää. Tietojemme mukaan tämä työ on ensimmäinen, jossa käytetään puoliohjattua oppimista sekä ekvivalenttia regularisaatiota nivelrikon yhteydessä.

Avainsanat: nivelrikko, syväoppiminen, polvi, ohjattu oppiminen, puoliohjattu oppiminen

Foreword

When I completed my MSc in Computer Engineering back in 2003, I considered myself done with my studies and entered the IT industry workforce. Fortunately, I changed my mind, and in 2018 I started a Master's program in Health Science. These three years of study have been challenging yet given me so much, through good times and bad.

This project started over a year ago is now finally complete. I would like to thank Dr. Aleksei Tiulpin for all support and guidance he has provided over the course of this project. Working with the excellent KNEEL method as my starting point, it was easy to do my implementation. Special thanks for the resources Dr. Tiulpin organized for this project.

I want to also thank co-supervisor Huy Hoang Nguyen for support and guidance. With both of your help, not only was I able to complete my thesis, but I also learned to change my way of thinking.

I also extend special thanks to the Faculty of Medicine for their Master Thesis grant and to the MIPT team for all their support.

Oulu, 24.05.2021

Raimo Niemelä

Abbreviations and Symbols used

AAM	Active Appearance Models
AI	Artificial Intelligence
ANN	Artificial Neural Network
ASM	Active Shape Model
CLM	Constrained Local Model
CT	computed tomography
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
FP	False Positive
GT	Ground truth
KL	Kellgren-Lawrence
ML	Machine learning
MRI	magnetic resonance image
OA	Osteoarthritis
OAI	Osteoarthritis Initiative
PCA	Principal component analysis
PS	Pictorial Structures
ROI	Region of Interest
RFRV	Random Forest Regression Voting
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machine
TP	True Positive

Table of contents

1 Introduction

Osteoarthritis (OA) is a common musculoskeletal disorder that affects millions of people regardless of race, sex, and gender around the world [24]. Of all the joints in the body, the knee is the heaviest weight-bearing joint and a highly prevalent location for the disease [29]. It is characterized by the breakdown of knee joint cartilage, the appearance of osteophytes, and the narrowing of joint space [29]. The disease causes knee stiffness and swelling, as well as aching after prolonged sitting or resting, gradually progressing to the point when patients have to undergo total knee replacement (TKR) surgeries to avoid full-fledged physical disability [37]. Unfortunately, effective cures for knee OA are not available yet. As the treatment for knee OA and TKR surgeries bring massive burdens in personal and societal levels, detecting knee OA at early stages to slow its progression is needed. In the diagnostic process for knee OA, radiography is the first imaging-based diagnostic tool [6, 19]. Radiography, known for its affordability and convenience, can capture dense tissues such as bones, which are informative enough for assessing the knee OA severity. One of the most important features that one can extract from radiographs is key points around the tibia and femur with respect to the perspective in the radiograph, from which joint space width could be derived [33]. However, annotating knee landmarks is costly in terms of time and budget. Therefore, computer aided programs have been developed to automate the task [62, 40, 41].

Anatomical landmark localization is challenging because structures are often similar [52]. Traditionally, the task was based on statistical methods, but recently more deep learning (DL)-based approaches have been proposed to improve its performance considerably [39, 69]. In particular, the convolutional neural networks (CNN) are suitable for landmark localization. Recently hourglass CNN has been successfully used for landmark detection [62], cell instance tracking and segmentation [53], segmentation of brain tumor [4] and human pose estimation [47]. They allow to capture information across all scales and are therefore good option for landmark localization. Original hourglass CNN predicts heatmaps and is memory demanding [61]. In the knee OA domain, the KNEEL method [62] is the one, which uses hourglass CNN and predicts directly landmarks points. The method utilizes pre-training and is accurate also on unseen data. Compared to the current state-of-the-art, KNEEL method has better generalization performance [62].

Although KNEEL achieved remarkable performance in knee OA, it lacked thorough investigation of hyperparameters, and did not explore the data requirements. In this work, the author first improved the baseline performance by conducting more sophisticated fine-tuning strategies that enable training with a much more efficient amount of annotated samples. Additionally, I propose to incorporate equivariant regularization [31] into the KNEEL approach in such a way that we can leverage unlabeled data and enable training KNEEL in a semi-supervised learning (SSL) setting.

Data used in the preparation of this thesis were obtained and analyzed from the controlled access datasets distributed from the Osteoarthritis Initiative (OAI)[1], a data repository housed within the NIMH Data Archive (NDA). OAI is a collaborative informatics system created by the National Institute of Mental Health and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS) to provide a worldwide resource to quicken the pace of biomarker identification, scientific investigation and OA drug development.

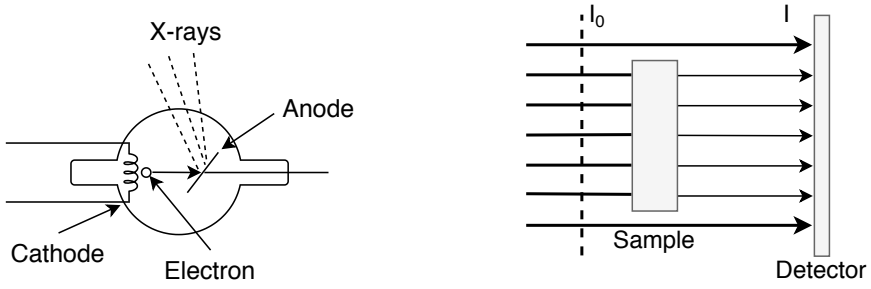
[1] https://nda.nih.gov/oai/

2 Background

2.1 Radiographic Image

In the field of medical imaging, a radiographic image (plain radiograph) is a 2-D image derived by sending X-rays through a certain human body part. As human organs and body parts are made of tissues with a diversity of densities that absorb radiation differently, the radiograph can reveal the underlying structure of the imaged body part, which is helpful for doctors to diagnose various diseases without performing surgery.

X-rays were discovered by Wilhelm Röntgen in 1895, and are electromagnetic radiation waves with a wavelength of $6.0 \cdot 10^{-12} - 1.5 \cdot 10^{-8}$ meters [45]. Figure 1.a depicts the structure of an X-ray tube. To generate X-rays, one passes electrons through a vacuum from the cathode to the anode. Moving electrons collide with other electrons in the anode, generating a large amount of energy from the interaction. Most of this energy is heat and must be removed. Only 1 % of this energy is X-radiation [45].



(a) Structure of X-ray tube (b) X-ray principle

Figure 1: X-ray tube and principle description

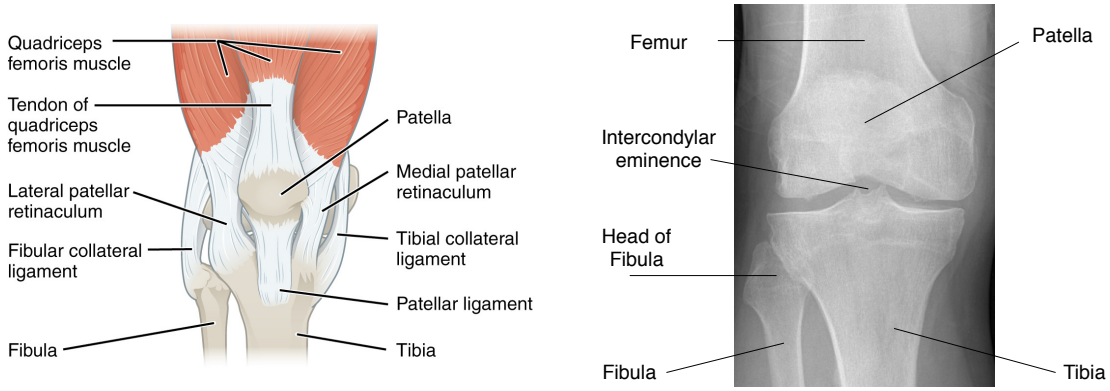
The X-rays are released from an X-ray tube with intensity I_0. An object absorbs electromagnetic radiation and intensity I is passed through the object (tissue) to the detector (see Figure 1.b). The X-ray image is a reflection of the intensities that passed through the object [1].

2.2 Knee Radiograph

The knee joint, the largest joint in the human body, articulates the femur, tibia and patella [5] (illustrated in Figure 2.a). In addition to the bones, there are soft tissues in the knee that are essential to the joint function: muscles, ligaments, and tendons. The muscles allow movement and keep the knee stable. The ligaments connect bones to each other, while the tendons connect muscles to the bones. The articular cartilage

provides a low-friction sliding surface between the bones. Both the femur and tibia can have rolling and gliding motions, allowing flexion and extension of the leg [5].

A knee X-ray image (Figure 2.b) is captured in the standing position with fixed flexion [61], merely revealing hard tissues. The quality of a knee radiograph depends on the pose position of the patient, and an adequate level of exposure and the X-ray beam angle [9, 62]. With the potential for low quality imagery and the lack of soft tissue in the image, diagnosing diseases based on knee radiographs is challenging and requires practitioners and radiologists to have intensive training courses to master the skill [62].



(a) Structure of the knee joint. [2].

(b) X-ray taken from knee [3].

Figure 2: Illustrated structure of the knee joint and X-ray of knee joint

2.3 Knee Landmark Localization

Anatomical landmark is a specific point in an image, which corresponds to an actual location in the body. Landmark localization aims to find these points with some automated method. It is an important part in many computer vision applications, such as expression understanding, face recognition [42] and eye tracking [51]. It can also be used with medical applications such as body composition [32], back pain problem analyses [16] and osteoarthritis analyses [63, 48].

Under a certain beam angle, knee joint space (KJS) appears as a gap between the tibia and femur in a knee radiograph [9]. KJS is one of the most important characteristics used to assess the loss of cartilage in knee OA diagnosis [9]. Fortunately, we can extract it from knee landmarks that are pre-defined key points around projected tibia and femur regions in the image. In total, there are 16 landmarks in a knee radiograph that are

[2]"Anterior view of right knee" by OpenStax is used under a CC BY 4.0 Licence. https://openstax.org/details/books/anatomy-and-physiology

[3]Data used in the preparation of this thesis were obtained and analyzed from the controlled access datasets distributed from the Osteoarthritis Initiative (OAI),https://nda.nih.gov/oai/

used for training, as depicted in Figure 3. The objective of knee landmark localization is to produce a list of key points $(x_i, y_i, t_i)_{i=1..N}$ describing positions (x_i, y_i) in the image coordinate and their type t_i from an input knee radiograph.

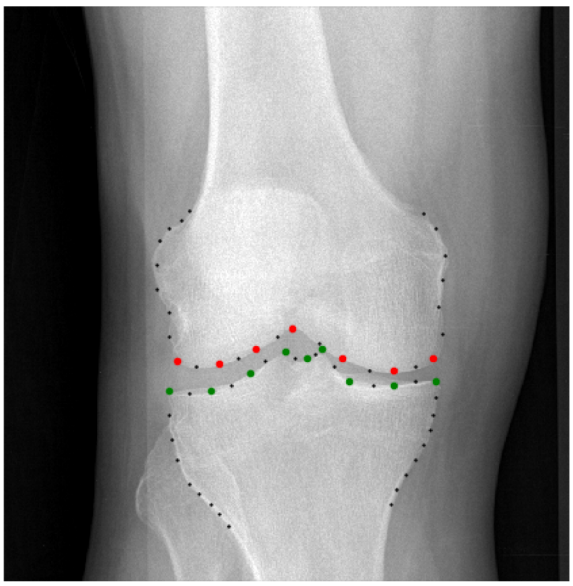


Figure 3: Knee landmarks in the X-ray image [4]. Tibia landmarks used for training marked in green and femur landmarks in red. The rest of the landmarks have not been used for training (marked in black). The landmarks were localized using Bone-Finder[5]

2.4 Machine Learning

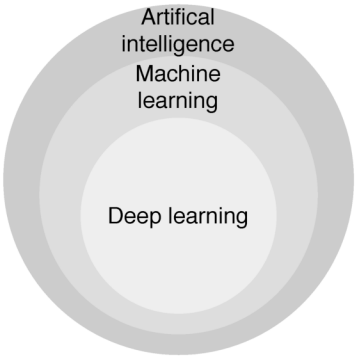


Figure 4: Artificial intelligence, Machine and Deep learning

Machine learning (ML) is a subfield of Artificial Intelligence (AI) and covers the deep learning (DL) field as shown in Figure 4. In ML, a machine learns to perform a certain

[4] https://nda.nih.gov/oai/
[5] http://bone-finder.com/

task based on given data. The objective is to have a learned machine that can generalize the task on unseen inputs instead of purely remembering the given data.

ML can be divided into three main categories: supervised learning (SL), semi-supervised learning (SSL) and unsupervised learning (UL) [59]. The primary difference among them is the involvement of ground truth data. Typically, there are two types of data such as annotated data (input samples x_i with corresponding targets y_i) and unannotated data (only input samples x_i) [2]. In supervised learning, all data must be annotated before it can be used; therefore, the input data are a set L of input-output pairs. Different from SL, SSL- and UL-based methods involve another set U of unannotated samples. While SSL has both L and U, UL only has U.

In the next subsections, we present in detail SL, SSL and the artificial neural network (ANN), one of the most well-known techniques in ML.

2.4.1 Supervised Learning

Given a set of annotated samples $D = \{(x_i, y_i)\}_{i=1..N}$, let f_θ be an arbitrary parametric differentiable function with parameter vector θ. The objective of SL is to find θ such that the empirical loss $\mathcal{L}_l(f_\theta, L)$ is minimized. Formally, we define θ such that

$$\arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i)), \tag{1}$$

where $\mathcal{L}$ is a loss function.

2.4.2 Semi-Supervised Learning

Let $D = \{(x_i, y_i)\}_{i=1..N}$ and $U = \{u_j\}_{j=1..M}$ be annotated and unannotated sets, respectively. Let f_θ denote an arbitrary function with parameters θ. SSL aims to utilize both annotated and unannotated samples to gain better generalization. The objective is to minimize a linear combination of empirical supervised and unsupervised losses, $\mathcal{L}_l$ and $\mathcal{L}_u$ respectively, with respect to f_θ. Formally, we find θ such that

$$\arg\min_\theta \left[\frac{w_l}{N} \sum_{i=1}^{N} \mathcal{L}_l(f_\theta(x_i), y_i) + \frac{w_u}{M} \sum_{j=1}^{M} \mathcal{L}_u(f_\theta(u_j)) \right], \tag{2}$$

where w_l and w_u are the coefficients of the supervised and unsupervised terms, respectively.

2.4.3 Artificial Neural Network

An artificial neural network, or simply neural network (NN) is a common architecture with simulated neurons, originally designed to mimic behaviors of the human brain [34]. The simplest form of an NN is a perceptron representing a linear binary classifier f that allows mapping input vector $\mathbf{x} = (1, x_1, ..., x_d)^T$ to $y \in \{0, 1\}$. Let $\theta = (\theta_0, \theta_1, ..., \theta_d)^T$ denote parameters of f, and we have:

$$f_\theta(\mathbf{x}) = \begin{cases} 1 & \text{if } \theta^T\mathbf{x} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We can generalize equation (3) as

$$f_\theta(\mathbf{x}) = \varphi(\theta^T\mathbf{x}), \tag{4}$$

where $\varphi(\cdot)$ is an activation (e.g. sigmoid of tanh) function (discussed in Section 2.5.4). While a sigmoid function transforms the prediction into $[0, 1]$, a tanh function maps it to $[-1, 1]$. Figure 5a illustrates a perceptron with an input layer $\mathbf{x}$ connected to output layer $\mathbf{y}$ by weights θ and a nonlinear function φ.

The complexity of an NN can be increased by adding hidden layers between the input and output layers. Figure 5b presents a neural network with two hidden layers in which all neurons are connected to neurons in the next level. Here, the architecture has 1 node in the output layer, and the circles can be any linear or nonlinear function (e.g. identity, sigmoid, tanh, etc.).

At the start of the training, the weights are initialized randomly [23, 27]. During the training, weights are optimized so that differences between the ground truth and predicted results are minimized. Training is carried out in phases: first inputs are converted to outputs by forwarding the inputs through the neural network, and the output is calculated. Then, the loss function is used to evaluate the error between actual output and the predicted output. Finally, a gradient of the loss is computed, and later used to optimize the network's parameters.

2.5 Deep Learning

2.5.1 Convolutional Neural Network

Convolution operator and convolutional kernel[6] are the fundamental elements in convolutional neural networks (CNN) [35]. Convolution slides a convolutional kernel, learn-

[6] Strictly speaking, convolutional networks perform cross-correlation.

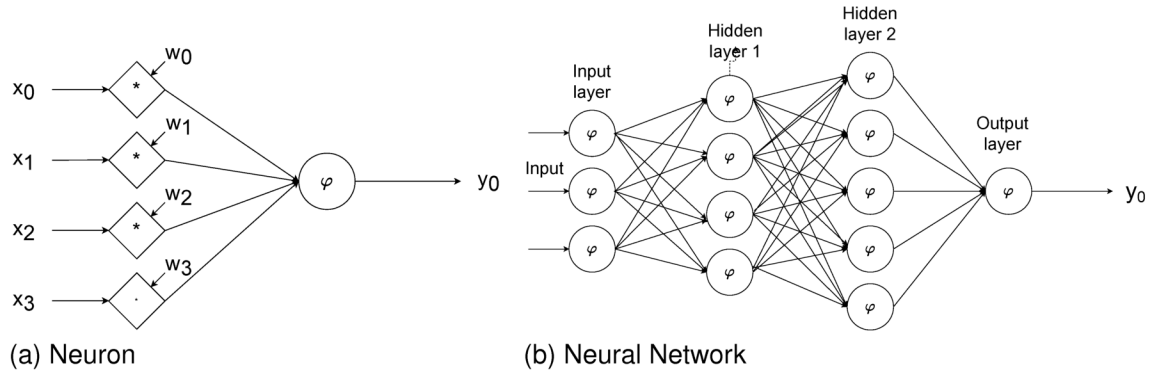


Figure 5: Neuron and neural network

able weights, throughout an input image or a feature map to produce an output feature map whose elements are a linear combination between the kernel's weights and input layer's values. A simple example of convolution and kernel can be found in Figure 6. Each of the convolution layers specializes in identifying a certain type of feature. Convolutional layers close to the inputs can learn low-level features, while layers near the outputs learn more abstract aspects. Figure 7 depicts an example of a CNN architecture with 2 convolutional layers. CNN-based architectures are widely used to automatically learn features from data in computer vision (CV), automatic sound recognition (ASR) and natural language processing (NLP) [54].

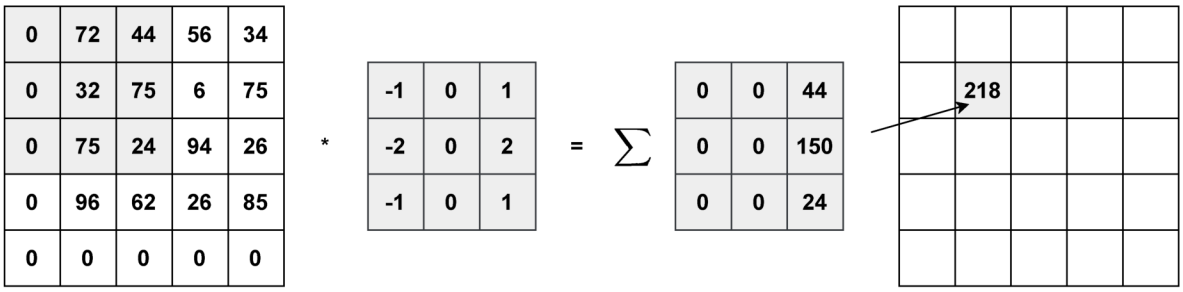


Figure 6: Kernel is moved over the input matrix cell by cell

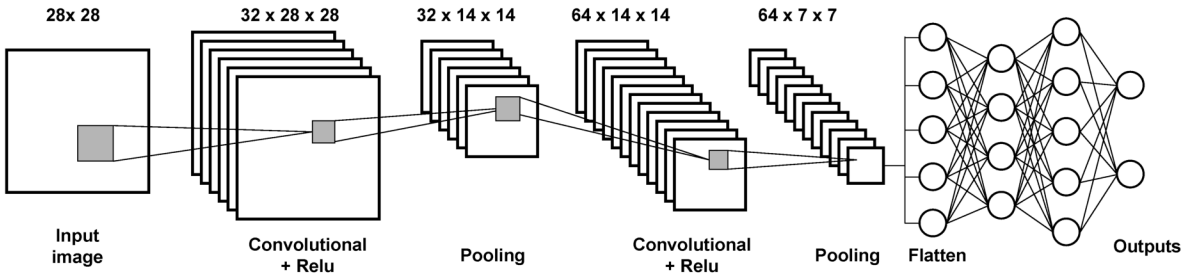


Figure 7: Example structure of Convolutional neural network

2.5.2 Pooling Layer

The pooling layer is a very common operator in DL used to downsample feature maps by a certain factor. Figure 8 presents two widely used pooling layers: average pooling and max pooling [35]. Similar to convolutional layers, the pooling layer scans a window with a certain size through input feature maps to aggregate its values (e.g., by averaging or getting maximums). In the literature, while max pooling is often used between middle layers to downscale the size of feature maps, average pooling is applied to the feature map of the last convolutional layer to "flatten" it.

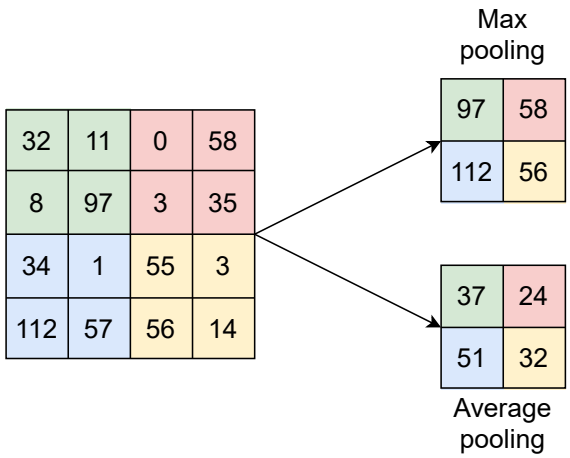


Figure 8: Differences between average and max pooling

2.5.3 Fully-Connected Layer

In this layer, each input is connected to all neurons. This is a standard structure in all kinds of neural networks. Computationally, this can become expensive when the numbers of inputs grow large. Therefore this layer is used for specific purposes such as classification. In CNNs, this layer is usually located after the convolutional and pooling layers as presented in Figure 7.

2.5.4 Activation Functions

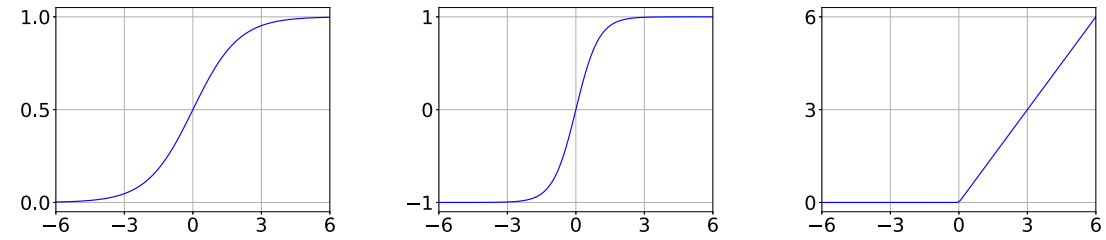
An activation function is either a linear or nonlinear function that is applied to a certain node of a neural network. In practice, most of the activation functions are non-linear [50]. Some of the widely used activation functions are Sigmoid, Tanh, and rectified linear unit (ReLU) functions [50] (illustrated in Figure 9). Sigmoid is defined

as:

$$y = \frac{1}{1 + e^{-x}}. \tag{5}$$

While the output values of Sigmoid are in the range $[0, 1]$, the Tanh function yields values in $[-1, 1]$ as its definition is

$$y = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{6}$$



(a) Sigmoid function (b) Tanh function (c) ReLU function

Figure 9: Activation functions

The ReLU function, the most heavily used activation nowadays in convolutional neural networks, is mathematically represented as follows [60]:

$$ReLU(x) = max(0, x). \tag{7}$$

Softmax is another important activation function, specialized for use in multi-class classification or attention-based neural networks. Given $x = [x_1, \ldots, x_n]^t$, the Softmax function, denoted by σ, such that

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}, \tag{8}$$

outputs an n-element vector equal to 1 such that $\sum_{i=1}^{n} \sigma(x_i) = 1$.

2.5.5 Loss Function

In ML, we aim to minimize or maximize certain objective function. When an objective is minimized, it is called a loss function, and describes the difference between the prediction and ground truth targets. Formally, given an input data $D = \{(x_i, y_i\})_{i=1}^{n}$, a loss function is a map $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}$, and the objective is to find parameters θ of function f such that:

$$\min_{\theta} \mathcal{L}(f_\theta, D), \tag{9}$$

Here we present some common and relevant loss functions in the landmark localization domain, such as $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_{wing}$ losses defined as follows,

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{n} |f_\theta(x_i) - y_i|, \tag{10}$$

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{n} (f_\theta(x_i) - y_i)^2, \tag{11}$$

$$\mathcal{L}_{wing}(y_i, f_\theta(x_i)) = \begin{cases} w \log\left(1 + \frac{1}{\epsilon}|y_i - f_\theta(x_i)|\right) & |y_i - f_\theta(x_i)| < w \\ |y_i - f_\theta(x_i)| - C & \text{otherwise.} \end{cases} \tag{12}$$

While $\mathcal{L}_1$ measures mean absolute error (MAE), $\mathcal{L}_2$ measures the mean squared error (MSE). Wing loss is closely related to $\mathcal{L}_1$ with constant C and nonlinear part $(-w, w)$ [62, 21].

2.6 Data Augmentation

Data augmentation is a popular technique that helps to enrich understanding of the data manifold and improve generalization of the model [30, 56]. In the scope of this thesis, we are interested in 2D point and image augmentations. A transformation T is a function that maps a point $\mathbf{x} = [x_1, x_2]^\intercal$ or an image I to another point $\mathbf{x}' = T(\mathbf{x})$ or image $I' = T(I)$. As the transformation can be represented by matrices, we alternatively write $\mathbf{x}' = T\mathbf{x}$ or $I' = TI$.

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{13}$$

As transformations are matrices, multiple transformations can be applied in the form of transformation composition.

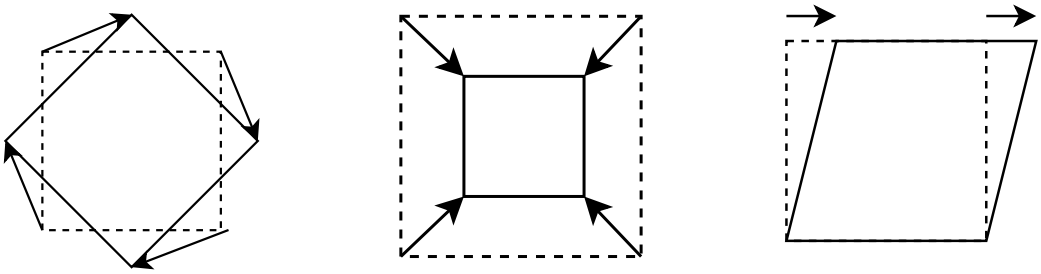
$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = T_1 T_2 T_3 \mathbf{x} = T_1 T_2 T_3 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{14}$$

2.6.1 Rotation

Rotation is the movement of a shape around a specific point. Figure 10a presents the principle of rotation transformation. Rotation is done through the fixed point with a fixed angle. Mathematically rotation can be defined as follows [46]:

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} cos\alpha & -sin\alpha \\ sin\alpha & cos\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{15}$$

where x_1 and x_2 are original coordinates, α rotation angle and x_1' and x_2' coordinates after rotation transform.



(a) rotation (b) scaling (c) shear

Figure 10: Rotation, scaling and shear transformations

2.6.2 Scaling

Scaling is a linear transformation that makes the size of the object smaller or bigger. Figure 10b presents the principle of scaling transformation. The size of the object increases or decreases by a scale factor s. The shape will remain the same but the size will be different after transformation. Mathematically, the scaling can be represented as follows [22]:

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{16}$$

where x_1 and x_2 are original coordinates, and x_1' and x_2' coordinates after transformation and s scaling factor.

2.6.3 Shear

Shear is a transformation in which the shape of the object changes and becomes slanted. Shear transformation can be done in either x or y axis. Figure 10c presents the principle of shear transformation in x axis, where the coordinates in the x direction are shifted to shear direction. Shear can be defined mathematically as follows [22]:

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 1 & sh_{x1} \\ sh_{x2} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \tag{17}$$

where x_1 and x_2 are coordinates of original image, sh_x and sh_y are shear factors, and x_1' and x_2' are coordinates after transformation.

2.6.4 MixUp

Given two arbitrary inputs x_1 and x_2, and λ sampled from $Beta(\alpha, \alpha)$, MixUp performs the convex combination

$$x' = \lambda x_1 + (1 - \lambda) x_2. \tag{18}$$

MixUp can be seen as a data-agnostic augmentation method [67] that can generate out-of-manifold samples [26, 62, 48]. Figure 11 presents the principle of MixUp.

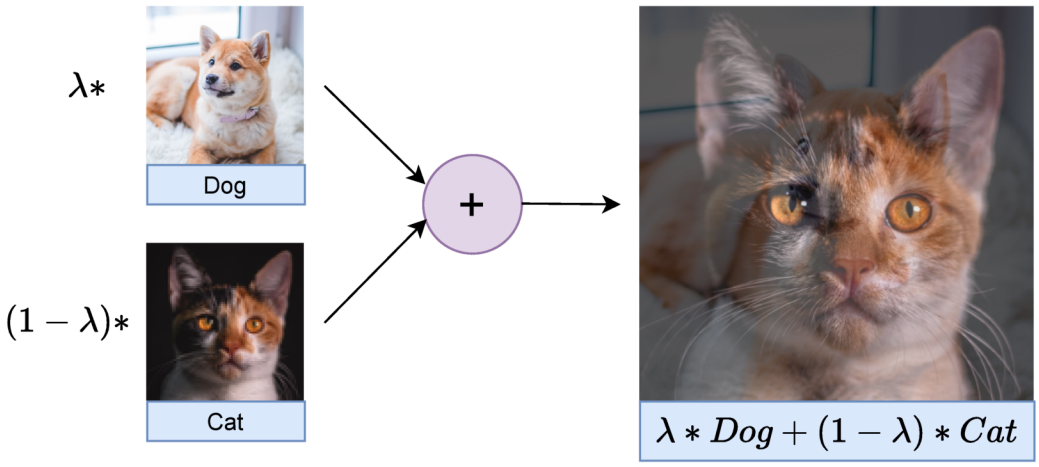


Figure 11: An illustration of Mixup applied on to a pair of images with λ of 0.3.

3 Related Work

3.1 Shape-based Models

To localize vertebrae landmarks, Lecron *et al.* [38] proposed to extract Scale-Invariant Feature Transform descriptors (SIFT) [44] of points of interest in an X-ray image. Then, they utilized multi-class Support Vector Machine (SVM) to classify whether SIFT descriptors are associated with a vertebra or not. Potesil *et al.* [20] utilized. Tim Cootes applied Active Shape Model (ASM) [11] to localize the boundary of knee cartilage of each MR imaging slice. Then, Cootes improved the performance of the knee cartilage localization by replacing ASM by Active Appearance Model (AAM) [12]. In that study, a shape model of knee cartilage was created using labeled landmarks. Then, a grey-level texture model was added to the shape model, which collaborated more information available than in [11].

Cootes *et al.* [15] improved AAM method by adding local feature templates and presented the idea of Constrained Local Models (CLM). In the study, facial landmarks were localized. The model shape and texture models are built in the same way as in AAM [15, 14]. In AAM the whole object is modelled, but in [15] the model has a set of local feature templates [15].

Criminisi *et al.* [13] used multi-class random regression forests to localize anatomical structures. They used this method for locating the bounding box within computer tomography (CT) scans, but not for individual landmarks. In a random forest, the decision is made using tree predictors [7]. Each tree makes individual decisions and the final decision is made by voting. The class which has the most votes wins. Random forests can be used also with regression.

Lindner *et al.* [41] combined RFRV and CLM fitting. The best position was voted for every feature point by an CLM and RFRV combination. They tested the method with facial images and hand radiographs (37 landmark points). They showed that the combination is fast and accurate for point detection. Later Lindner *et al.* [40] used this combination to localize landmark points in knee radiograpsh and this is generally considered as state-of-the-art [62].

3.2 Machine Learning for Landmark Localization

In recent years, a few trials have been done with deep neural networks and medical data. Zhang *et al.* [69] localized landmarks from brain T1 weighted magnetic resonance (MR) images and prostate CT images. They combined the random forest technique with CNN to localize anatomical landmarks. Emad *et al.* [18] used CNN with seven

layers to detect the left ventricle from an MRI. They did not try to find actual landmark points, but a bounding box enclosing the left ventricle. Auber *et al.* [3] combined a deep neural network (DNN) with a statistical shape model (SSM) to detect the spine and the pelvis from plain X-ray images.

In general, anatomical landmark localization using CNN is based on heatmaps. Song *et al.* [58] predicted landmarks from orthodontic X-rays. First they extracted ROIs from X-ray and detected the landmarks from ROI patches. They trained CNN for every single landmark and predicted totally the location of 19 landmarks.

Nibali *et al.* [49] added Diferentiable Spatial to Numerical Transform (DSNT) layer to the CNN and predicted coordinates directly. Later, Yeh *et al.* [66] predicted 45 landmarks directly from a whole-spine lateral radiograph with modified Cascaded pyramid Network(CPN). They used heatmaps to predict the probable locations of landmarks, but used DSNT to calculate landmark predictions.

Softa *et al.* [57] used Fully Convolutional Neural Network (FCN) [43] to predict keypoint locations from ultrasound images. They used regression maps and used the center of mass of the regression maps as the location estimate. Davison *et al.* [17] located landmarks using U-Net [55]. First they located reference points from pelvic radiogpraph and then predicted the offset to the reference point. Finally target point was voted using the predicted target points.

Trigeorgis *et al.* [64] combined a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN) to predict face alignment. Feng *et al.* [21] presented a new loss function, Wing loss and combined it with CNN for facial landmark localization. Zhang *et al.* [68] presented accurate face alignment method using Coarse-to-Fine Auto-Encoder Networks (CFAN).

Payer *et al.* [52] used regression based CNN with heatmap for landmark localization from hand, skull and spine. Full-resolution heatmaps are problematic with medical data, because the size of the medical image is typically large [62]. But decreasing the resolution may may affect the accuracy of landmark localization [62]. Tiulpin *et al.* [62] used hourglass CNNs to find an intermediate solution between the landmarks and the direct predictions. Earlier hourglass CNNs have been used to estimate human pose, because hourglass CNN works well on low resolution images [47].

4 Methods and Materials

4.1 Hourglass Network

Hourglass network is an encoder-decoder network, in which feature maps of its input image are spatially shrunk through its encoder, and expanded back to its original spatial shape through its decoder, which forms a hourglass-like shape. Although originally proposed to estimate human poses appearing at different scales in generic images [47, 25], the architecture later showed its strength in the medical imaging domain [62].

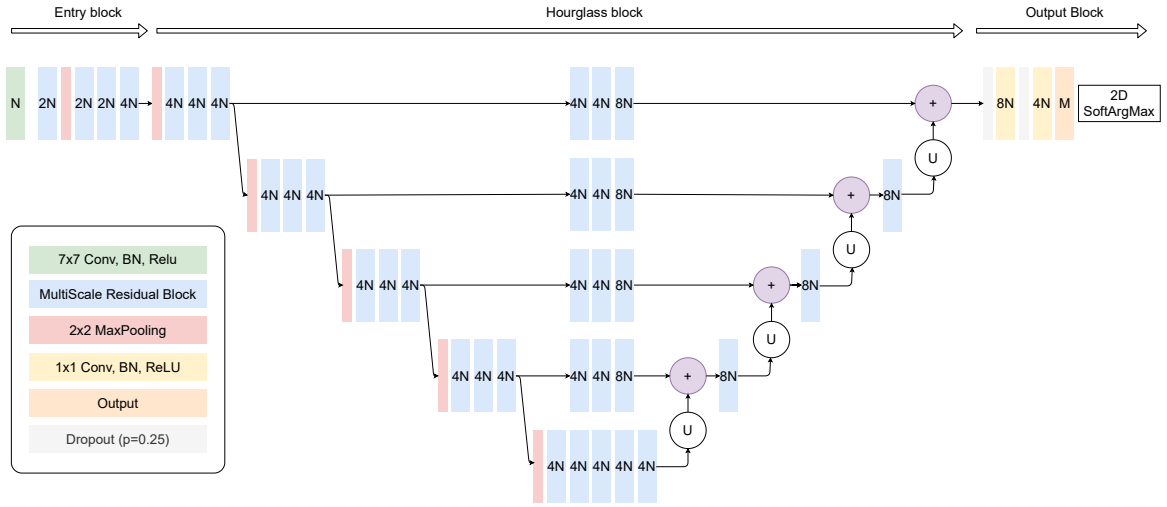


Figure 12: Description of the hourglass network use in this study. The depth of the hourglass block is 4, N is the number of initial feature maps, and M is the number of output heatmaps.

In this study, we design a hourglass network to learn to predict knee joint landmarks in an end-to-end manner. Our hourglass network has three types of blocks: entry, hourglass, and output as shown in 12. First, the entry block located at the beginning is a shallow sub-network consisting of a convolutional block (in green) and residual modules (in blue) separated by a max-pooling layer (in red). Second, the hourglass block, the main component, comprises an encoder and an decoder sub-networks. While we use max-pooling layers to perform sub-samplings in the encoder, we interpolate to up-sampling feature maps in the decoder. At each step in the encoder, we make a summation between feature maps up-sampled from the previous step and projections of the corresponding step in the decoder via residual blocks. Finally, feature maps derived from the hourglass block are passed through the output block, which includes dropout layers (in gray), 1×1 convolutional blocks (in yellow), and, especially, a 2D SoftArgMax layer.

4.2 Equivariant Regularization

As mentioned earlier, image augmentations enrich the model's understanding of the data manifold, help it to be less vulnerable to overfitting. However, without any constraints, we cannot ensure that the model will make rotated predicions if an input image is rotated. To explicitly enforce it, we utilize equivariant regularization [10]. Figure 13 illustrates the idea of equivariant regularization. Formally, we minimize

$$\frac{1}{N}\sum_{i=1}^{N} \|f_\theta(T\mathbf{x}_i) - Tf_\theta(\mathbf{x}_i)\|_1 \tag{19}$$

where N is the number of samples, f_θ is the neural network, and T is a random transformation. As no labels are needed, $\mathbf{x_i}$'s can be either labeled or unlabeled data. Eq. (19) is equivalent to enforcing consistency between the transformed landmarks and landmarks calculated from the transformed image [31] as follows

$$f_\theta(T\mathbf{x_i}) \approx Tf_\theta(\mathbf{x}_i), \forall i = 1..N. \tag{20}$$

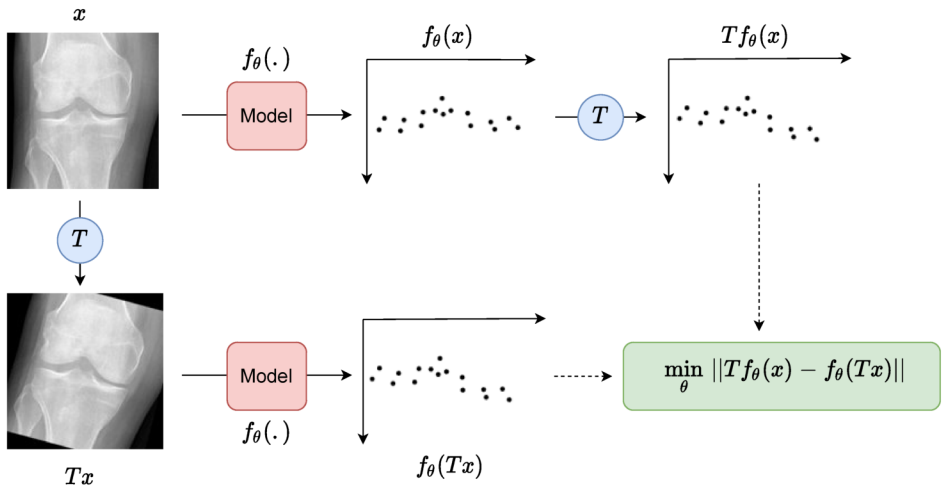


Figure 13: Image[7] is transformed during training and landmarks are predicted. Landmarks are predicted from the original image and landmarks are transformed.

4.3 Our Method

Based on the KNEEL method [62], we propose a robust and data-efficient method – called KNEEL+ – for knee joint landmark localization. Figure 14 presents the rela-

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

tionship between KNEEL and KNEEL+. As such, we fine-tune the hyperparameters of KNEEL to produce a more optimized version of KNEEL. The details of the hyperparameters can be found in Section 5.1.1. Moreover, we utilize the equivariant regularization to help our model to observe more data and be robust to perspective changes simultaneously. Specifically, we involve rotation, scaling, and shear transformations. Furthermore, we extended our method to semi-supervised learning setting by including unlabeled samples in the training process.

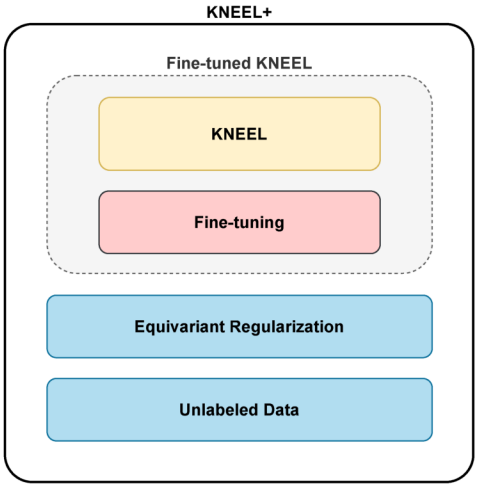


Figure 14: Development process of KNEEL+

The network was trained using the landmarks shown in Figure 15. Landmarks 0, 8, 9, 15 were used for evaluation. The landmarks are the same as in the original KNEEL [62].

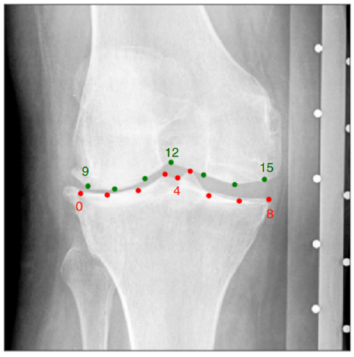


Figure 15: A knee sample of 16 landmarks are aimed to be localized [8]

[8] https://arxiv.org/abs/1907.12237

5 Experiments

5.1 Experimental Details

5.1.1 Implementation

We implemented our method based on the KNEEL [62][9], deep-pipeline[10] and SOLT[11] repositories. Following the KNEEL code, we used the Python programming language and Pytorch 1.1.0 in our implementation. Thanks to the supercomputer cluster Puhti with a total of 682 central processing unit (CPU) and 80 graphical processing unit (GPU) nodes (4 NVidia Volta V100 processors GPUs each), we could run our experiments in parallel. Each training setting was sent to the cluster's job list and waited for its turn to run automatically. We trained all models with a batch size of 16. We utilized the Adam optimizer [36] to train all settings with a fixed batch size of 16 and a learning rate of $1e-3$. Following [62], we used the wing loss [21]. Usually the learning rate is reduced during training using adaptive learning rate methods or predefined learning rate schedules. In this experiment, predefined learning rate schedules were used. Different from [62], which used predefined learning rate ($1e-3$), our implementation uses scheduler that drops the learning rate at 50 and 150 epochs. Besides, we trained all the settings in 300 epochs, instead of 100 as in [62]. All other experiments were done with learning rate drop at 50 and 150 epochs. Also, several training times were tried. The values we used can be found from Table 1. The transformations with equivariant regularization are presented in Table 3.

Table 1: Training parameters

Parameter	Value
Initial learning rate	$1e-3$
Learning drop schedule	[50, 150]
Learning rate after drop	$1e-4$, $1e-5$
Epochs	100, 150, 200, 250, 300

5.1.2 Data Preparation

Our experiments were conducted on the dataset extracted from the Osteoarthritis Initiative (OAI) [12] cohort. OAI is a public dataset and contains clinical and imaging data

[9] https://github.com/MIPT-Oulu/KNEEL
[10] https://github.com/lext/deep-pipeline
[11] https://github.com/MIPT-Oulu/solt.git
[12] https://nda.nih.gov/oai/

from an eleven-year cohort study with 4,796 patients [28]. All the patients at the moment of the recruitment were 45-79 years old and had risk developing or have developed knee OA. In addition to the baseline examination, there are 13 follow-up visits from 12 to 132 months and knee radiographs imaged at 18, 30, 48, 60, 84, 108, 120, and 132-month follow-ups. The dataset for training and model selection was the same as in the original KNEEL (totally 748 knee joints, approximately 150 per KL grade).

The annotations for our dataset are from [62]. To investigate the effect of annotation quantity on the performance of our method, we prepared 7 annotated data settings in which we randomly selected N/k annotated samples, where N is the number of all OAI data and $k = 1..7$ (see Table 2). Both low-cost and high-cost models were trained with the same data settings. To make the results of those settings compatible, we kept the validation set of each setting the same (149 samples).

Table 2: Amount of labeled data used with training

Labeled data	N	$\frac{1}{2}$ N	$\frac{1}{3}$ N	$\frac{1}{4}$ N	$\frac{1}{5}$ N	$\frac{1}{6}$ N	$\frac{1}{7}$ N
#Training	599	299	200	150	120	100	86
#Validation	149	149	149	149	149	149	149

In addition, we had a data setting for our SSL approach. We used 750 knee radiographs from the 12-month follow-up as unannotated samples. All unlabeled training data for high-cost training were created automatically using a script. First, the ROI was predicted using a low-cost model, and high-cost training data was extracted using the predicted ROI coordinates.

Table 3: Transformations used in the experiments

Transformation	**Configuration**					
	1	2	3	4	5	6
Rotation	-5°- 5°	-10°- 10°	-20°- 20°	-40°- 40°	-	-
Scale	0.8 - 1	0.9 -1	0.95 -1	1-1.05	1 - 1.1	1 - 1.2
Shear vertical	0 - 0.05	0 - 0.1	-	-	-	-
Shear horizontal	0 - 0.05	0 - 0.1	-	-	-	-

5.1.3 Experimental Protocol

Labeled data were split into 5 folds using K-fold cross-validation. The principle of K-fold cross validation is presented in Figure 16. In K-fold cross-validation, data are first split into k folds, and training is done with $k-1$ folds. In the first split, folds $0-3$ are

used for training, while fold 4 is used for validation. This is repeated until all folds are used for validation. This helps to use the data more efficiently.

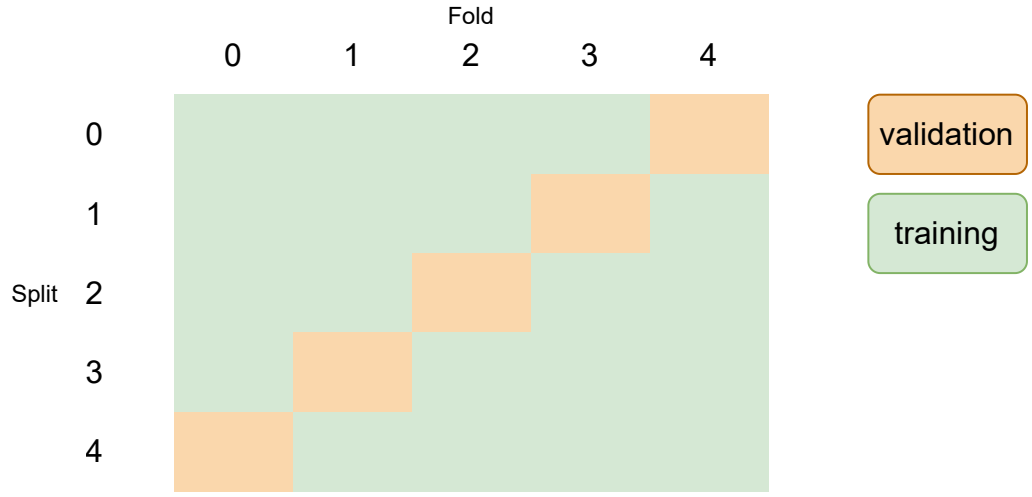


Figure 16: K-fold cross validation with k-value 5

In the first step, the training parameters were analyzed with different data amounts. The learning rate drop schedule and epoch count effect were analyzed. We had two approaches to initialize our model's weights. We either utilized the weights of a model that had been pre-trained with low-cost labels, or initialized them randomly. Low-cost models were trained to localize ROIs from the bilateral radiograph, while high-cost models used 16 femur and tibia anatomical landmark points. Both low-cost and high-cost training were performed with the same split.

Subsequently, we conducted an ablation study on image transformations for equivariant regularization. Firstly, we performed experiments with different rotations and amounts of labeled samples. Specifically, we rotated images with a range of degrees: $[-5, 5]$, $[-10, 10]$, $[-20, 20]$ or $[-40, 40]$. Because training takes a lot of time and computing resources were limited, the amount of labeled data used with equivariant regularization was selected after the first trials with rotation transformation were done. Rotation, scaling and shear-transformations were analyzed independently with labeled data amount N/3. The best transformations were selected (rotation angle in $[-5, 5]$ degrees, scale in $[1, 1.05]$, shear horizontal in $[0.0, 0.1]$) and combined into one.

Thirdly, it was investigated whether model could be further improved using unlabeled data. We kept the unlabeled a data amount the same (i.e., 748 samples) for all the experiments. We used previously selected transformations with equivariant requlariza-tion (rotation $[-5, 5]$, scale $[1, 1.05]$ and shear $[0.0, 0.1]$). All experiments were done with data amount mentioned in Table 2.

Finally, we compared the performances of KNEEL [62], the optimized KNEEL, KNEEL+,

and KNEEL+ with unlabeled data with different amounts of samples. Original KNEEL and optimized KNEEL were trained with and without pre-training (w/o low cost).

5.1.4 Metrics

Percentage of Correct Keypoints (PCK) [65] were calculated for all experiments. PCK is widely used with computer vision applications [70], [62], [8]. The distance between the predicted landmark and the ground truth was calculated. The percentage of predicted landmarks inside radius r is then presented (recall). Radius 1.0 mm, 1.5 mm, 2.0 mm and 2.5 mm were used with the experiments. The recall represents the proportion of landmarks correctly detected out of all landmarks within the current radius. Mathematically it can be defined as:

$$Recall = \frac{TP}{TP + FN}. \tag{21}$$

5.2 Results

5.2.1 Pre-training and Fine Tuning Training Parameters

Recall percentages for KNEEL and optimized KNEEL with different labeled data amounts and with different precision thresholds $[1.0, 1.5, 2.0, 2.5]$ mm can be found from Figure 17. It can be seen, that pre-training can be skipped only if full amount of the labeled data in use and threshold 2.5 mm can be accepted. Optimized KNEEL gives better results than original KNEEL. The results is better, even pre-training is not used with optimized KNEEL. The less labeled data in use, the bigger difference between the methods.

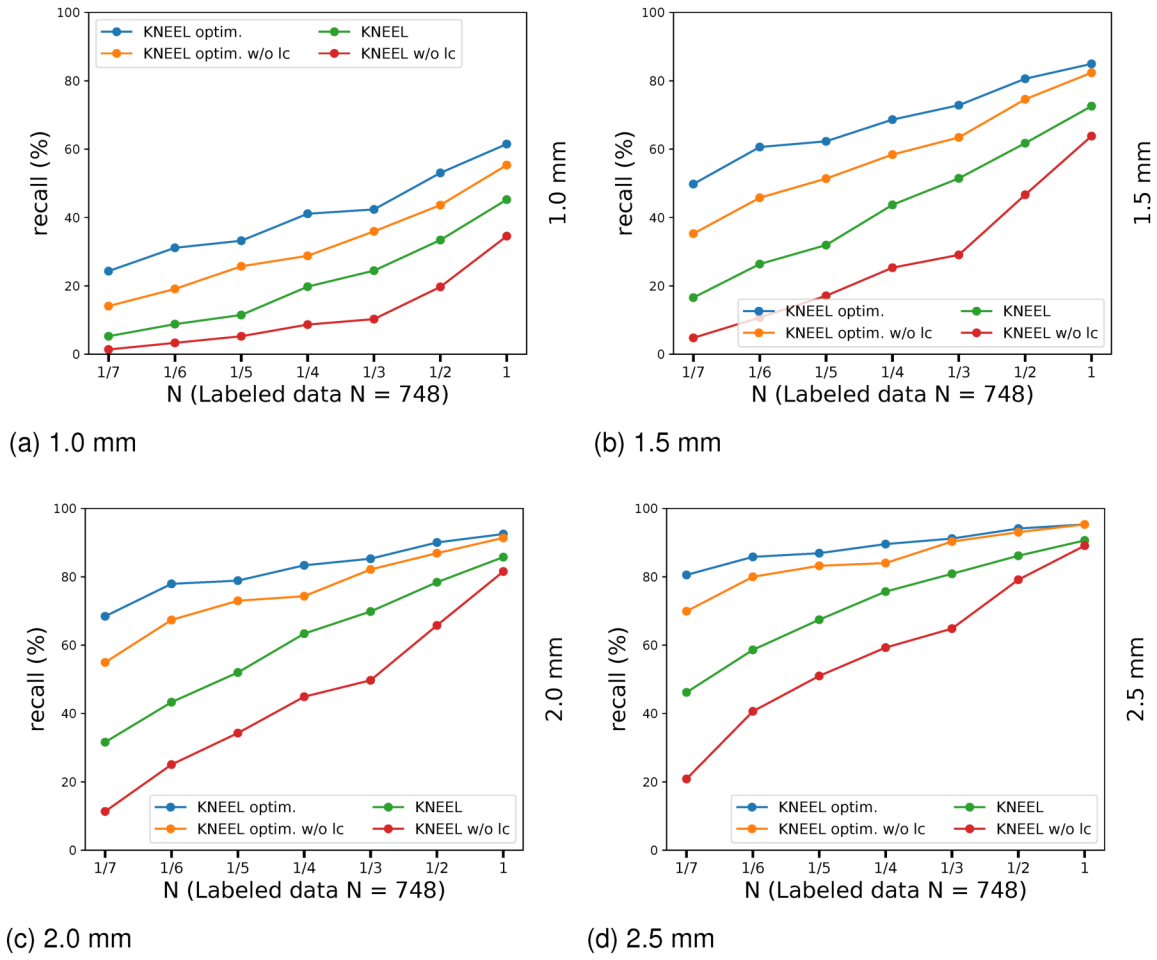


Figure 17: KNEEL / optimized KNEEL with and without pre-training.

5.2.2 Equivariant Regularization

Figure 18 presents the recall with different labeled data amounts with rotation transformation. It can be seen that with a labeled data amount $1/3N$ the accuracy varies a lot with radius 1.0 mm, and the labeled data amount is notable. Therefore, $1/3N$ amount of labeled data was selected for all transformatios for further investigation.

Secondly, the effect of scaling transformation was analyzed. Table 4 presents the effect of scaling transformation. It can be seen, that using scaling factor [1.0 - 1.05] the best performance can be achieved. Good results were also achieved with [0.9 - 1.0] with precision thresholds [1.5, 2.0] mm, but the former was chosen for its good accuracy with 1.0 mm threshold.

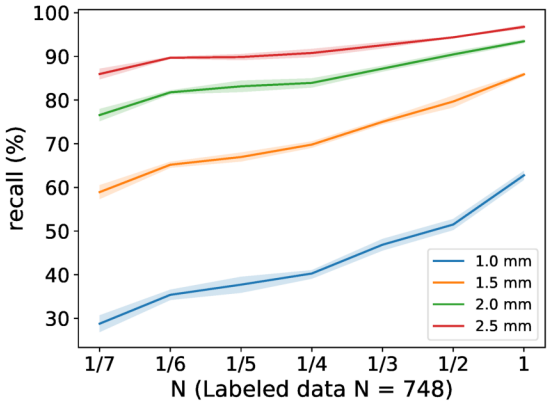


Figure 18: Data amount impact with equivariant regularization with angles [5,10,20,40]. Min and max values shaded. Average value drawn as a line.

Table 4: Scaling transformation effect to the landmark localization

Scale	# Labeled in train set	1 mm	1.5 mm	2 mm	2.5 mm	% out
0.8 - 1.0	$\frac{1}{3}$ N	50.00 ± 9.83	76.34 ± 6.43	87.83 ± 3.97	93.52 ± 1.61	0.40
0.9 - 1.0		49.67 ± 7.85	**79.75 ± 5.77**	**89.51 ± 4.44**	93.58 ± 2.65	0.80
0.95 - 1.0		49.33 ± 11.53	76.87 ± 5.48	87.83 ± 4.92	93.45 ± 2.46	0.67
1.0 - 1.05		**53.48 ± 11.34**	78.41 ± 4.44	**89.51 ± 4.06**	**93.98 ± 2.84**	0.80
1.0 - 1.1		51.00 ± 9.93	78.21 ± 6.43	88.84 ± 2.93	93.38 ± 2.55	0.27
1.0 - 1.2		52.74 ± 9.93	77.07 ± 4.63	89.30 ± 3.40	93.92 ± 1.42	0.40

The effect of shear transformation is shown in Table 5. Trials were done by shearing the image in both height and width directions. Shear values $[0.0, 0.05]$ and $[0.0, 0.1]$ were selected randomly in both directions. The best result was achieved with horizontal shear value $[0.0.\ 0.1]$ in 1.0 mm and 1.5 mm. Finally, the previously presented transformations were combined into one. The results are presented in Table 6. It can be seen, that combined transformations yield better results than the individual ones.

Table 5: Shear transformation affect on landmark localization

Shear	# Labeled in train set	1 mm	1.5 mm	2 mm	2.5 mm	% out
x: 0.0 - 0.05	$\frac{1}{3}$ N	53.81 ± 7.09	79.48 ± 4.82	88.97 ± 2.93	94.12 ± 1.89	0.67
x: 0.0 - 0.1		**54.88 ± 9.55**	**80.01 ± 6.33**	90.11 ± 3.78	94.18 ± 2.74	0.94
x: 0.0 - 0.05, y: 0.0 - 0.05		49.80 ± 10.87	77.27 ± 7.00	89.24 ± 4.06	93.78 ± 1.99	0.27
x: 0.0 - 0.1, y: 0.0 - 0.1		52.61 ± 10.30	78.61 ± 5.67	89.64 ± 4.06	94.12 ± 2.65	0.67
y: 0.0 - 0.05		52.74 ± 8.98	78.21 ± 5.67	89.84 ± 4.16	**94.85 ± 1.23**	0.53
y: 0.0 - 0.1		53.48 ± 11.15	79.48 ± 5.39	**90.17 ± 2.55**	94.12 ± 1.13	0.67

Table 6: Combined transform gives better results.

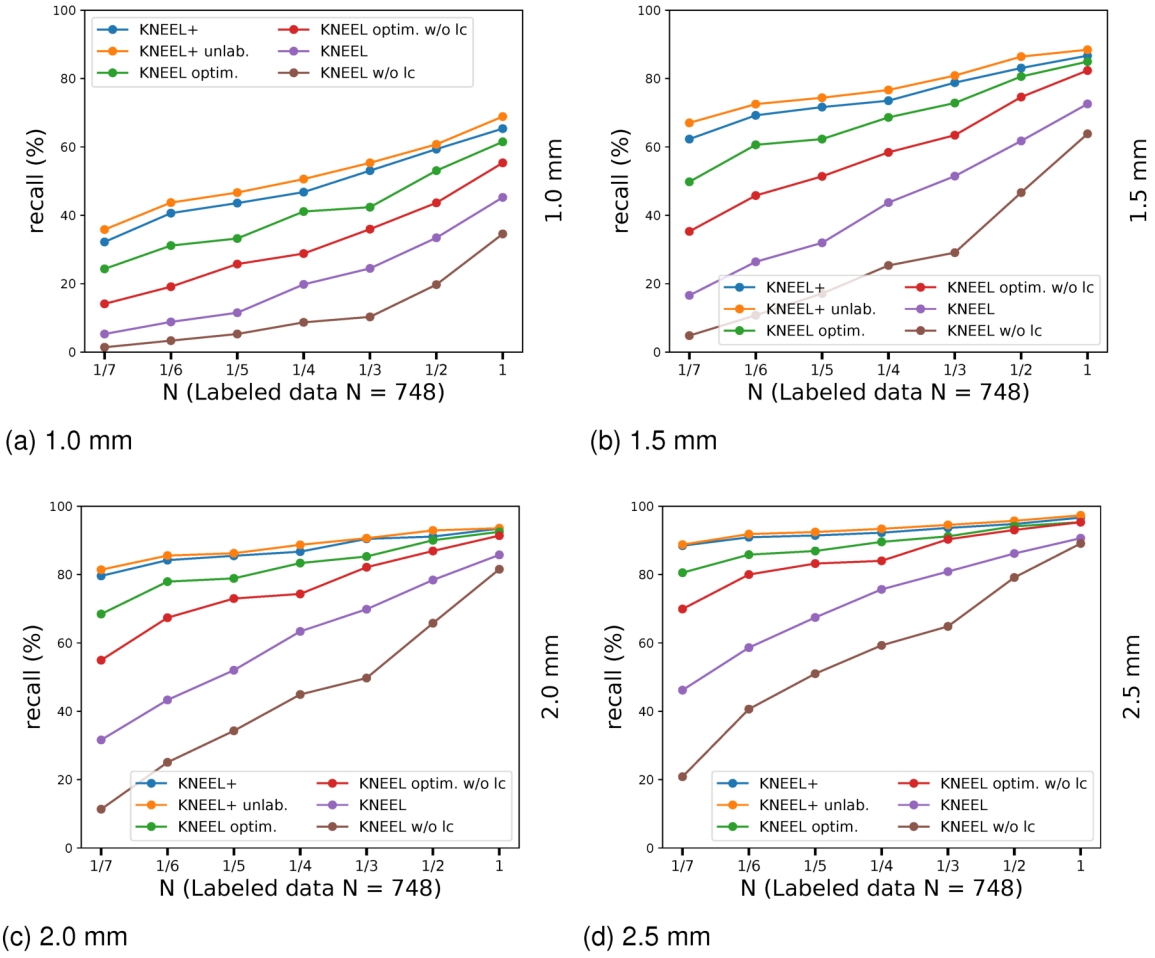
Transform	# Labeled in train set	1 mm	1.5 mm	2 mm	2.5 mm	% out
angle -5 - 5	$\frac{1}{3}$ N	50.94 ± 9.08	76.54 ± 5.77	88.84 ± 3.88	93.58 ± 3.03	0.53
scale 1.0 - 1.05		53.48 ± 11.34	78.41 ± 4.44	89.51 ± 4.06	93.98 ± 2.84	0.80
shear x: 0.0 - 0.1		53.81 ± 7.09	79.48 ± 4.82	88.97 ± 2.93	94.12 ± 1.89	0.67
all combined		**53.88 ± 12.48**	**80.61 ± 7.75**	**89.91 ± 4.06**	**94.92 ± 1.89**	0.53

5.3 Label Efficiency

Figure 19 presents, how the amount of labeled data affects to the performance. Comparison is done for KNEEL, optimized KNEEL and KNEEL+. The more labeled data used for training, the better the algorithm performs. KNEEL+ performs well even with the small amount of labeled data, if precision threshold 2.0 mm is accepted. This can also be seen with the optimized KNEEL. With the original KNEEL, performance decreases rapidly as the amount of labeled data decreases.

5.4 Unlabeled Data Utilization

A comparison between labeled and unlabeled data is presented in Table 7. The comparison also includes the original KNEEL and the optimized KNEEL. It can be seen from Table 7, that KNEEL+ trained with the unlabeled data gives best results. This can be seen with all amounts of labeled data. Performance increases by several percentage points especially when a small difference between GT and predicted landmark is required. By combining KNEEL+ with unlabeled data, less labeled data is needed as can be seen from Figure 22. Training the network with one-fifth of the labeled data gives results as good as original implementation with the full data amount. Figure 20 presents difference between KNEEL and KNEEL+ with same amount of the labeled data, KNEEL+ performs much better. This can also be seen with the labeled data amount 1/7N as presented in 21.



(a) 1.0 mm

(b) 1.5 mm

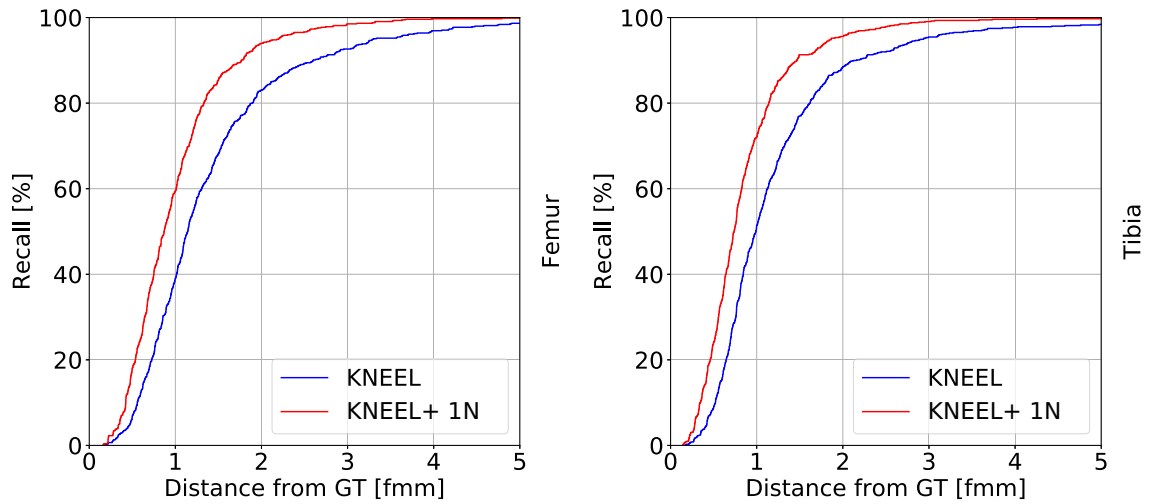
(c) 2.0 mm

(d) 2.5 mm

Figure 19: KNEEL / optimized KNEEL / KNEEL+ with and without pre-training on low-cost model.

Table 7: A comparison between labeled and unlabeled data

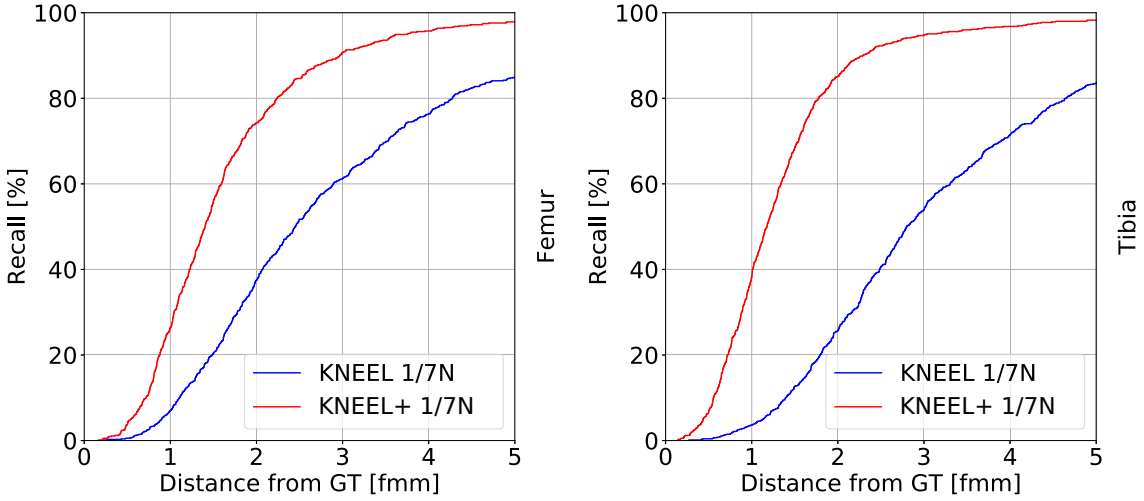
Method	# Labeled in train set	1 mm	1.5 mm	2 mm	2.5 mm	% out
KNEEL	N	45.25 ± 8.79	72.59 ± 6.24	85.76 ± 3.88	90.64 ± 1.89	0.80
KNEEL optimized		61.50 ± 10.21	84.96 ± 4.06	92.51 ± 1.70	95.32 ± 0.57	0.13
KNEEL+		65.37 ± 7.37	86.70 ± 5.20	93.45 ± 2.84	96.66 ± 0.76	0.40
KNEEL+ unlabeled 750		**68.89 ± 8.80**	**88.44 ± 2.83**	**93.56 ± 1.89**	**97.33 ± 0.94**	0.00
KNEEL	$\frac{1}{2}$ N	33.42 ± 5.10	61.76 ± 4.35	78.41 ± 2.93	86.16 ± 1.42	2.27
KNEEL optimized		53.07 ± 7.94	80.61 ± 5.86	90.04 ± 2.93	94.12 ± 2.08	1.07
KNEEL+		59.36 ± 10.78	83.09 ± 5.20	91.11 ± 2.17	94.79 ± 1.89	0.27
KNEEL+ unlabeled 750		**60.76 ± 11.06**	**86.43 ± 5.39**	**92.91 ± 3.40**	**95.72 ± 1.70**	0.27
KNEEL	$\frac{1}{3}$ N	24.47 ± 3.03	51.47 ± 4.73	69.85 ± 2.55	80.88 ± 2.08	3.74
KNEEL optimized		42.38 ± 6.81	72.86 ± 6.43	85.29 ± 4.35	91.18 ± 2.65	0.80
KNEEL+		53.07 ± 9.83	78.81 ± 7.28	90.44 ± 4.06	93.65 ± 3.12	0.53
KNEEL+ unlabeled 750		**55.35 ± 12.10**	**80.88 ± 6.62**	**90.64 ± 2.84**	**94.52 ± 2.46**	0.13
KNEEL	$\frac{1}{4}$ N	19.79 ± 1.13	43.72 ± 0.95	63.37 ± 1.13	75.67 ± 0.38	5.48
KNEEL optimized		41.11 ± 8.04	68.65 ± 5.39	83.36 ± 4.63	89.57 ± 2.65	1.74
KNEEL+		46.79 ± 9.08	73.53 ± 9.26	86.70 ± 4.06	92.25 ± 2.08	1.34
KNEEL+ unlabeled 750		**50.60 ± 8.98**	**76.67 ± 6.14**	**88.70 ± 3.12**	**93.38 ± 2.17**	0.94
KNEEL	$\frac{1}{5}$ N	11.50 ± 0.76	31.95 ± 1.13	52.01 ± 2.65	67.45 ± 1.99	7.35
KNEEL optimized		33.22 ± 6.71	62.30 ± 7.18	78.88 ± 6.81	86.90 ± 3.78	3.21
KNEEL+		43.58 ± 8.89	71.66 ± 8.70	85.49 ± 5.77	91.44 ± 1.89	1.07
KNEEL+ unlabeled 750		**46.66 ± 9.08**	**74.40 ± 8.04**	**86.23 ± 5.10**	**92.45 ± 3.12**	1.07
KNEEL	$\frac{1}{6}$ N	8.82 ± 0.57	26.40 ± 0.28	43.32 ± 0.38	58.62 ± 0.47	9.09
KNEEL optimized		31.15 ± 2.84	60.63 ± 1.61	77.94 ± 2.08	85.83 ± 1.70	2.81
KNEEL+		40.64 ± 5.10	69.25 ± 6.62	84.22 ± 4.73	90.91 ± 1.70	2.41
KNEEL+ unlabeled 750		**43.72 ± 6.43**	**72.53 ± 6.14**	**85.56 ± 4.16**	**91.84 ± 1.32**	0.94
KNEEL	$\frac{1}{7}$ N	5.28 ± 2.36	16.58 ± 5.48	31.62 ± 8.22	46.19 ± 7.85	11.50
KNEEL optimized		24.33 ± 1.32	49.80 ± 1.42	68.45 ± 0.95	80.55 ± 1.99	5.21
KNEEL+		32.22 ± 8.32	62.30 ± 8.70	79.61 ± 7.66	88.44 ± 5.39	1.74
KNEEL+ unlabeled 750		**35.83 ± 7.00**	**67.05 ± 6.14**	**81.42 ± 4.73**	**88.77 ± 2.84**	2.14



(a) Recall with femur landmarks

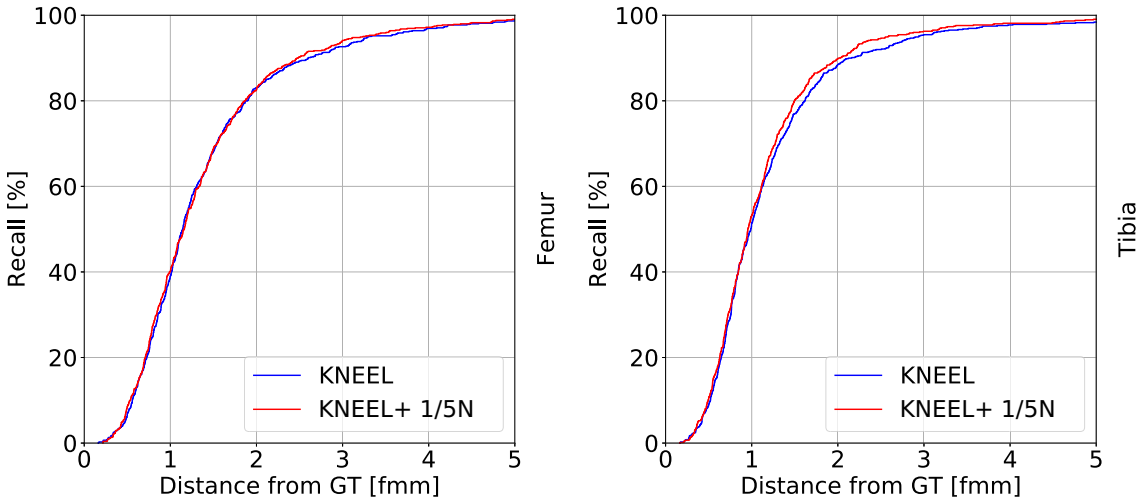
(b) Recall with tibia landmarks

Figure 20: KNEEL+ performs better with same amount of labeled data (+ Unlabeled data 748 pcs)



(a) Recall with femur landmarks (b) Recall with tibia landmarks

Figure 21: Comparison between KNEEL and KNEEL+ with data amount $1/7N$ (+ unlabeled data 748 pcs)



(a) Recall with femur landmarks (b) Recall with tibia landmarks

Figure 22: Less data is needed with KNEEL+. Training the network with one-fifth of the data gives results as good as original implementation with the full data amount (N)

6 Discussion

The aim of this work was to investigate whether the previously developed KNEEL method can be improved. In addition, the effect of the training data amount on the performance had to be analyzed, as well as the effect of the unlabeled data.

Longer training with a modified learning rate drop schedule (optimized KNEEL) gives better results than the original KNEEL. It gives better results without a low-cost model than the original KNEEL, but training with a low-training model will further improve results. Therefore, pre-training is needed for landmark localization. It can be skipped only if an error greater than 2.5 mm between the ground truth and the predicted landmark is accepted and the labeled data amount is bigger than $1/3N$.

With the labeled data amount N and optimized KNEEL, 61.50% of the predicted landmarks were inside 1 mm radius, while without modifications only 45.25% (KNEEL paper) were inside the same radius. Pre-training is needed for all amounts of labeled data if a small error between the GT landmark and predicted landmark is requested. Only if the full amount of labeled data are available and 2.5 mm error is accepted, the pre-training can be skipped. The difference is significant between the original and improved KNEEL: within 1.0 mm radius, the difference between recall is 15 - 20 percentage points in all data amounts. Within the larger radius, the difference is bigger when the labeled data amount is smaller. The conclusion is that learning rate drop schedule and longer training time significantly increase performance.

Secondly, it was investigated whether the method could be improved using equivariant regularization. In the study, several transformations were tried. The end result is that combining transformations gives better results than individual transformations. Using scale, rotation and shear transformation together, 65.37% of the predicted landmarks were inside 1 mm radius (vs. KNEEL 45.24%). Equivariant regularization has an advantage, because it can be used with unlabeled data. In the previous steps, trained networks were utilized with unlabeled data. The end result is that the network can learn more from unlabeled data. The difference is noticeable with the full amount of training data (KNEEL+ with unlabeled data 68.89% vs. KNEEL+ 65.37%). Using a smaller amount of the training data, the difference is still significant. Using 20% of the original training data, the improved method predicts 46.66% landmarks inside 1 mm radius, while the original KNEEL method predicts only 11.50%.

Figure 20 presents that with the same amount of labeled data ($1N$) KNEEL+ performs better. While KNEEL predicts about 40% of the femur landmarks inside 1.0 mm radius, KNEEL+ can predict about 60% of the landmarks correctly. The results for tibia landmarks are 50% and 70% respectively.

Using a smaller amount of data, the difference is even more significant. This can be

seen in Figure 21, where the labeled data amount for training is $1/7N$. With KNEEL+ 80% of the predicted tibia landmarks are closer than 2 mm from GT, while under 30% of the predicted landmarks are closer than 2.0 mm for GT with the original KNEEL. With femur landmarks, the difference is smaller, but still significant. Under 5% of predicted tibia landmarks are inside 1 mm radius with the original implementation, while KNEEL+ predicts about 40% landmarks inside the same radius. The difference between KNEEL and KNEEL+ is noticeable with all amounts of labeled data. Using unlabeled data with labeled data, the performance increases a few percentages with a small radius. When the radius between GT and the predicted landmark increases, the unlabeled data does not bring much benefit.

Figure 22 presents the recall curves for the original KNEEL and the improved KNEEL+. The amount of labeled data for KNEEL+ is $1/5N$ while KNEEL has the full amount of labeled data(1N). Both methods predict 40% of femur landmarks inside 1.0 mm radius and 50% of tibia marks inside 1.0 mm radius. The results are similar, but KNEEL+ needs only 20% of the labeled data compared to KNEEL. More detailed results can be found in Appendix 3. Without unlabeled data, KNEEL+ needs about 25% of the labeled data, while KNEEL+ with the unlabeled data needs 20% of the labeled data to achieve the same results as the original KNEEL.

This work has multiple limitations. Firstly, our method has been used only for the knee joints. In the future, more research should be done with other types of X-rays as well. The training parameters should be redefined for new types of X-rays. Secondly, A static validation data amount was used in all experiments (149 pcs), but it would have been better to use a certain percentage share. The experiments were executed with new validation data amounts (20%), and the results can be found from Appendix 3 and Appendix 4. It seems that even better results can be achieved with small amounts of training data.

7 Conclusions

This thesis focused on landmark localization using deep learning. In practice, different training parameters were tested and the effect on prediction accuracy studied. At the same time, the effect of training data amount were investigated. Finally, we looked at whether equivariant regularization can give better results with unlabeled data.

From the results it can be seen that longer training leads to better results. Using equivariant regularization, the results are significantly better than with the original implementation. It was seen that when three transforms were combined, better results were achieved with equivariant regularization. In addition, performance can be improved by training the network using unlabeled data. However, the author notes that the major benefit from unlabeled data comes only when the number of training samples is very limited. When the amount of training samples is large, it is sufficient to use only the equivariant regularizartion.

Finally, the improved method KNEEL+ gives significantly better results than the original KNEEL. With KNEEL+ the same results can be achieved using only 20% of the labeled data compared to the original implementation. This matters, because annotation of the medical images is expensive and time consuming. The downside is that training takes more time than it does with the original KNEEL.

To conclude this work, the presented results demonstrate how one can improve upon the state-of-the-art in knee anatomical landmark localization. The main application of the developed methodology is in the knee OA research, where it is essential to analyze large patient cohorts, such as OAI[13]. The author believes that the developed methodology is general can easily be adapted in the fields outside osteoarthritis.

Acknowledgments

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.

[13] `https://nda.nih.gov/oai/`

References

[1] Y Abdallah and R Yousef. “Augmentation of X-rays images using pixel intensity values adjustments”. In: *International Journal of Science and Research (IJSR)* 4.2 (2015), pp. 2425–2430.

[2] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[3] Benjamin Aubert et al. “Automatic spine and pelvis detection in frontal X-rays using deep neural networks for patch displacement learning”. In: *2016 ieee 13th international symposium on biomedical imaging (isbi)*. IEEE. 2016, pp. 1426–1429.

[4] Eze Benson et al. “Deep hourglass for brain tumor segmentation”. In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 419–428.

[5] J. Gordon Betts et al. *Anatomy and Physiology*. Ed. by J. Gordon Betts.

[6] Hillary J Braun and Garry E Gold. “Diagnosis of osteoarthritis: imaging”. In: *Bone* 51.2 (2012), pp. 278–288.

[7] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[8] Yu-Wei Chao et al. “Forecasting human dynamics from static images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 548–556.

[9] HC Charles et al. *Optimization of the fixed-flexion knee radiograph*. 2007.

[10] Taco Cohen and Max Welling. “Group equivariant convolutional networks”. In: *International conference on machine learning*. 2016, pp. 2990–2999.

[11] Timothy F. Cootes. “An Introduction to Active Shape Models”. In: 1992.

[12] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. “Active appearance models”. In: *European conference on computer vision*. Springer. 1998, pp. 484–498.

[13] Antonio Criminisi et al. “Regression forests for efficient anatomy detection and localization in CT studies”. In: *International MICCAI Workshop on Medical Computer Vision*. Springer. 2010, pp. 106–117.

[14] David Cristinacce and Tim Cootes. “Automatic feature localisation with constrained local models”. In: *Pattern Recognition* 41.10 (2008), pp. 3054–3067.

[15] David Cristinacce and Timothy F Cootes. “Feature detection and tracking with constrained local models.” In: *Bmvc*. Vol. 1. 2. Citeseer. 2006, p. 3.

[16] Dimitrios Damopoulos, Ben Glocker, and Guoyan Zheng. “Automatic localization of the lumbar vertebral landmarks in CT images with context features”. In: *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer. 2017, pp. 59–71.

[17] Adrian K Davison et al. “Landmark localisation in radiographs using weighted heatmap displacement voting”. In: *International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer. 2018, pp. 73–85.

[18] Omar Emad, Inas A Yassine, and Ahmed S Fahmy. “Automatic localization of the left ventricle in cardiac MRI images using deep learning”. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2015, pp. 683–686.

[19] Parastu S Emrani et al. “Joint space narrowing and Kellgren–Lawrence progression in knee osteoarthritis: an analytic literature synthesis”. In: *Osteoarthritis and Cartilage* 16.8 (2008), pp. 873–882.

[20] Pedro F Felzenszwalb and Daniel P Huttenlocher. “Pictorial structures for object recognition”. In: *International journal of computer vision* 61.1 (2005), pp. 55–79.

[21] Zhen-Hua Feng et al. “Wing loss for robust facial landmark localisation with convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2235–2245.

[22] James D Foley et al. *Computer graphics: principles and practice*. Vol. 12110. Addison-Wesley Professional, 1996.

[23] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html.

[24] Sion Glyn-Jones et al. “Osteoarthritis”. In: *The Lancet* 386.9991 (2015), pp. 376–387.

[25] Chunsheng Guo, Wenlong Du, and Na Ying. “Multi-Scale Stacked Hourglass Network for Human Pose Estimation”. In: (2018).

[26] Hongyu Guo, Yongyi Mao, and Richong Zhang. “Mixup as locally linear out-of-manifold regularization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3714–3722.

[27] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[28] National Institutes of Health. *Osteoarthritis Initiative (OAI) Study Protocol*. https://nda.nih.gov/oai/study-details (accessed November 22, 2020). 2020.

[29] Behzad Heidari. “Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I”. In: *Caspian journal of internal medicine* 2.2 (2011), p. 205.

[30] Alex Hernández-García and Peter König. “Data augmentation instead of explicit regularization”. In: *arXiv preprint arXiv:1806.03852* (2018).

[31] Sina Honari et al. “Improving Landmark Localization with Semi-Supervised Learning”. In: *CoRR* abs/1709.01591 (2017). arXiv: 1709.01591. URL: http://arxiv.org/abs/1709.01591.

[32] Peijun Hu et al. “Automated characterization of body composition and frailty with clinically acquired CT”. In: *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer. 2017, pp. 25–35.

[33] Ravi Kant Jain, Abhishek Jain, and Pranav Mahajan. “Radiographic evaluation of knee joint space width using fixed flexion view in knees of Indian adults”. In: *International Journal of Research in Orthopaedics* 5.1 (2019), p. 38.

[34] Biing Hwang Juang. “Deep neural networks–a developmental perspective”. In: *APSIPA Transactions on Signal and Information Processing* 5 (2016).

[35] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. “A convolutional neural network for modelling sentences”. In: *arXiv preprint arXiv:1404.2188* (2014).

[36] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[37] Yu Ko et al. “Health-related quality of life after total knee replacement or unicompartmental knee arthroplasty in an urban asian population”. In: *Value in Health* 14.2 (2011), pp. 322–328.

[38] Fabian Lecron, Mohammed Benjelloun, and Saïd Mahmoudi. “Fully automatic vertebra detection in x-ray images based on multi-class SVM”. In: *Medical Imaging 2012: Image Processing*. Vol. 8314. International Society for Optics and Photonics. 2012, p. 83142D.

[39] Yuanwei Li et al. “Fast multiple landmark localisation using a patch-based iterative network”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 563–571.

[40] Claudia Lindner 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*. Springer. 2013, pp. 181–189.

[41] Claudia Lindner et al. “Robust and accurate shape model matching using random forest regression-voting”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2014), pp. 1862–1874.

[42] Wei Liu et al. “Deep Residual Equivariant Mapping for Multi-angle Face Recognition”. In: *Chinese Conference on Biometric Recognition*. Springer. 2019, pp. 145–154.

[43] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[44] David G Lowe. “Object recognition from local scale-invariant features”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[45] Harry E Martz et al. *X-ray Imaging: fundamentals, industrial techniques and applications*. CRC Press, 2016.

[46] James Nearing. “Mathematical tools for physics”. In: (2006).

[47] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation”. In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[48] Huy Hoang Nguyen et al. “Semixup: In-and Out-of-Manifold Regularization for Deep Semi-Supervised Knee Osteoarthritis Severity Grading From Plain Radiographs”. In: *IEEE Transactions on Medical Imaging* 39.12 (2020), pp. 4346–4356.

[49] Aiden Nibali et al. “Numerical coordinate regression with convolutional neural networks”. In: *arXiv preprint arXiv:1801.07372* (2018).

[50] Chigozie Nwankpa et al. “Activation functions: Comparison of trends in practice and research for deep learning”. In: *arXiv preprint arXiv:1811.03378* (2018).

[51] Seonwook Park et al. “Learning to find eye region landmarks for remote gaze estimation in unconstrained settings”. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 2018, pp. 1–10.

[52] Christian Payer et al. “Integrating spatial configuration into heatmap regression based CNNs for landmark localization”. In: *Medical image analysis* 54 (2019), pp. 207–219.

[53] Christian Payer et al. “Segmenting and tracking cell instances with cosine embeddings and recurrent hourglass networks”. In: *Medical image analysis* 57 (2019), pp. 106–119.

[54] Daniele Ravì et al. “Deep learning for health informatics”. In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 4–21.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[56] Patrice Simard et al. “Tangent prop-a formalism for specifying selected invariances in an adaptive network”. In: *Advances in neural information processing systems*. 1992, pp. 895–903.

[57] Michal Sofka et al. “Fully convolutional regression network for accurate detection of measurement points”. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2017, pp. 258–266.

[58] Yu Song et al. “Automatic cephalometric landmark detection on X-ray images using a deep-learning method”. In: *Applied Sciences* 10.7 (2020), p. 2547.

[59] Jost Tobias Springenberg. “Unsupervised and semi-supervised learning with categorical generative adversarial networks”. In: *arXiv preprint arXiv:1511.06390* (2015).

[60] Vivienne Sze et al. “Efficient processing of deep neural networks: A tutorial and survey”. In: *Proceedings of the IEEE* 105.12 (2017), pp. 2295–2329.

[61] A. (Aleksei) Tiulpin. *Deep learning for knee osteoarthritis diagnosis and progression prediction from plain radiographs and clinical data*. Ed. by S. (Simo) Saarakkala. Oulun yliopisto, 2020. URL: http://urn.fi/urn:isbn:9789526225524.

[62] Aleksei Tiulpin, Iaroslav Melekhov, and Simo Saarakkala. *KNEEL: Knee Anatomical Landmark Localization Using Hourglass Networks*. 2019. arXiv: 1907.12237 [cs.CV].

[63] Aleksei Tiulpin et al. *Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach*. 2017. arXiv: 1710.10589 [cs.CV].

[64] George Trigeorgis et al. “Mnemonic descent method: A recurrent process applied for end-to-end face alignment”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4177–4187.

[65] Yi Yang and Deva Ramanan. “Articulated human detection with flexible mixtures of parts”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.12 (2012), pp. 2878–2890.

[66] Yu-Cheng Yeh et al. “Deep learning approach for automatic landmark detection and alignment analysis in whole-spine lateral radiographs”. In: *Scientific reports* 11.1 (2021), pp. 1–15.

[67] Hongyi Zhang et al. “mixup: Beyond empirical risk minimization”. In: *arXiv preprint arXiv:1710.09412* (2017).

[68] Jie Zhang et al. “Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment”. In: *European conference on computer vision*. Springer. 2014, pp. 1–16.

[69] Jun Zhang, Mingxia Liu, and Dinggang Shen. “Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks”. In: *IEEE Transactions on Image Processing* 26.10 (2017), pp. 4753–4764.

[70] Christian Zimmermann and Thomas Brox. “Learning to estimate 3d hand pose from single rgb images”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4903–4911.

8 Appendices

Appendix 1.	Full comparison between KNEEL, optimized KNEEL, KNEEL+
Appendix 2.	Comparison between different rotation angles KNEEL+
Appendix 3.	Results with modified validation data amount
Appendix 4.	Full comparison between KNEEL, optimized KNEEL, KNEEL+ with new validation data amount

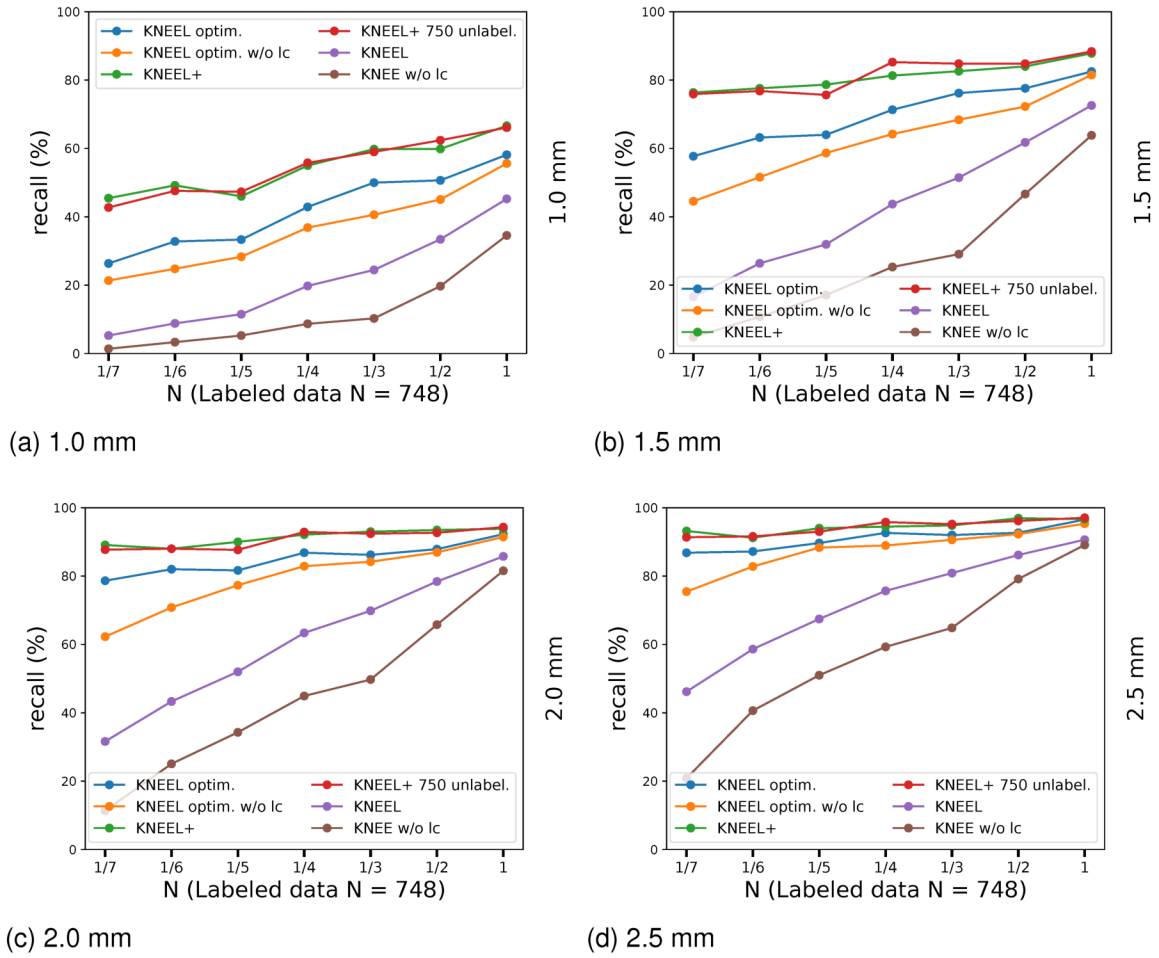
Appendix 1. Full comparison with / without pretraining on low-cost model

Method	# Labeled in train set	1 mm	1.5 mm	2 mm	2.5 mm	% out
KNEEL+		65.37 ± 7.37	86.70 ± 5.20	93.45 ± 2.84	96.66 ± 0.76	0.40
KNEEL+ unlab.		**68.89 ± 8.80**	**88.44 ± 2.83**	**93.56 ± 1.89**	**97.33 ± 0.94**	0.00
KNEEL optim.	N	61.50 ± 10.21	84.96 ± 4.06	92.51 ± 1.70	95.32 ± 0.57	0.13
KNEEL optim. w/o lc		55.35 ± 13.48	82.36 ± 6.74	91.39 ± 3.19	95.32 ± 1.66	0.84
KNEEL		45.25 ± 8.79	72.59 ± 6.24	85.76 ± 3.88	90.64 ± 1.89	0.80
KNEEL w/o lc		34.56 ± 2.74	63.84 ± 2.93	81.55 ± 1.51	89.10 ± 0.47	1.20
KNEEL+		59.36 ± 10.78	83.09 ± 5.20	91.11 ± 2.17	94.79 ± 1.89	0.27
KNEEL+ unlab.		**60.76 ± 11.06**	**86.43 ± 5.39**	**92.91 ± 3.40**	**95.72 ± 1.70**	0.27
KNEEL optim.	$\frac{1}{2}$ N	53.07 ± 7.94	80.61 ± 5.86	90.04 ± 2.93	94.12 ± 2.08	1.07
KNEEL optim. w/o lc		43.65 ± 12.76	74.60 ± 11.72	86.90 ± 7.00	93.05 ± 3.40	0.80
KNEEL		33.42 ± 5.10	61.76 ± 4.35	78.41 ± 2.93	86.16 ± 1.42	2.27
KNEEL w/o lc		19.72 ± 3.50	46.66 ± 3.59	65.78 ± 2.65	79.14 ± 1.51	3.07
KNEEL+		53.07 ± 9.83	78.81 ± 7.28	90.44 ± 4.06	93.65 ± 3.12	0.53
KNEEL+ unlab.		**55.35 ± 12.10**	**80.88 ± 6.62**	**90.64 ± 2.84**	**94.52 ± 2.46**	0.13
KNEEL optim.	$\frac{1}{3}$ N	42.38 ± 6.81	72.86 ± 6.43	85.29 ± 4.35	91.18 ± 2.65	0.80
KNEEL optim. w/o lc		35.96 ± 5.10	63.44 ± 7.66	82.15 ± 6.14	90.31 ± 3.69	1.47
KNEEL		24.47 ± 3.03	51.47 ± 4.73	69.85 ± 2.55	80.88 ± 2.08	3.74
KNEEL w/o lc		10.29 ± 1.51	29.08 ± 0.66	49.73 ± 0.95	64.84 ± 1.13	4.41
KNEEL+		46.79 ± 9.08	73.53 ± 9.26	86.70 ± 4.06	92.25 ± 2.08	1.34
KNEEL+ unlab.		**50.60 ± 8.98**	**76.67 ± 6.14**	**88.70 ± 3.12**	**93.38 ± 2.17**	0.94
KNEEL optim.	$\frac{1}{4}$ N	41.11 ± 8.04	68.65 ± 5.39	83.36 ± 4.63	89.57 ± 2.65	1.74
KNEEL optim. w/o lc		28.81 ± 2.93	58.42 ± 4.16	74.33 ± 3.59	84.02 ± 2.17	2.94
KNEEL		19.79 ± 1.13	43.72 ± 0.95	63.37 ± 1.13	75.67 ± 0.38	5.48
KNEEL w/o lc		8.69 ± 2.46	25.33 ± 1.80	44.92 ± 3.03	59.29 ± 1.42	7.89
KNEEL+		43.58 ± 8.89	71.66 ± 8.70	85.49 ± 5.77	91.44 ± 1.89	1.07
KNEEL+ unlab.		**46.66 ± 9.08**	**74.40 ± 8.04**	**86.23 ± 5.10**	**92.45 ± 3.12**	1.07
KNEEL optim.	$\frac{1}{5}$ N	33.22 ± 6.71	62.30 ± 7.18	78.88 ± 6.81	86.90 ± 3.78	3.21
KNEEL optim. w/o lc		25.74 ± 4.06	51.40 ± 5.20	72.99 ± 4.54	83.22 ± 4.06	3.21
KNEEL		11.50 ± 0.76	31.95 ± 1.13	52.01 ± 2.65	67.45 ± 1.99	7.35
KNEEL w/o lc		5.28 ± 1.04	17.11 ± 1.13	34.29 ± 1.04	51.00 ± 2.93	7.89
KNEEL+		40.64 ± 5.10	69.25 ± 6.62	84.22 ± 4.73	90.91 ± 1.70	2.41
KNEEL+ unlab.		**43.72 ± 6.43**	**72.53 ± 6.14**	**85.56 ± 4.16**	**91.84 ± 1.32**	0.94
KNEEL optim.	$\frac{1}{6}$ N	31.15 ± 2.84	60.63 ± 1.61	77.94 ± 2.08	85.83 ± 1.70	2.81
KNEEL optim. w/o lc		19.12 ± 0.38	45.79 ± 1.04	67.38 ± 3.59	80.01 ± 3.31	3.21
KNEEL		8.82 ± 0.57	26.40 ± 0.28	43.32 ± 0.38	58.62 ± 0.47	9.09
KNEEL w/o lc		3.34 ± 1.51	10.76 ± 2.93	25.07 ± 2.74	40.64 ± 3.21	9.63
KNEEL+		32.22 ± 8.32	62.30 ± 8.70	79.61 ± 7.66	88.44 ± 5.39	1.74
KNEEL+ unlab.		**35.83 ± 7.00**	**67.05 ± 6.14**	**81.42 ± 4.73**	**88.77 ± 2.84**	2.14
KNEEL optim.	$\frac{1}{7}$ N	24.33 ± 1.32	49.80 ± 1.42	68.45 ± 0.95	80.55 ± 1.99	5.21
KNEEL optim. w/o lc		14.10 ± 0.47	35.29 ± 0.57	54.95 ± 1.89	69.92 ± 3.78	4.28
KNEEL		5.28 ± 2.36	16.58 ± 5.48	31.62 ± 8.22	46.19 ± 7.85	11.50
KNEEL w/o lc		1.40 ± 0.66	4.81 ± 2.65	11.36 ± 4.73	20.86 ± 5.29	22.33

Appendix 2. Comparison between different rotation angles KNEEL+

Max rotation	# Labeled in train set	1 mm	1.5 mm	2 mm	2.5 mm	% out
angle 0	N	66.58 ± 6.24	**87.63 ± 4.06**	93.52 ± 2.17	96.66 ± 0.19	0.27
angle 5		**67.18 ± 7.85**	86.97 ± 5.39	93.45 ± 2.65	96.52 ± 1.13	0.13
angle 10		65.91 ± 8.70	87.10 ± 5.39	93.85 ± 2.27	97.06 ± 1.51	0.00
angle 20		64.97 ± 8.13	86.23 ± 4.35	**94.05 ± 1.61**	**97.13 ± 0.66**	0.00
angle 40		64.44 ± 8.51	86.23 ± 5.86	93.52 ± 1.23	96.59 ± 1.04	0.00
angle 0	$\frac{1}{2}$ N	57.75 ± 8.13	**83.36 ± 5.01**	91.78 ± 2.93	95.19 ± 2.08	0.80
angle 5		**58.29 ± 10.59**	82.49 ± 5.86	91.24 ± 3.69	94.52 ± 2.46	0.40
angle 10		57.02 ± 10.49	81.62 ± 5.39	**91.84 ± 4.35**	**95.59 ± 2.65**	0.00
angle 20		55.68 ± 10.68	81.89 ± 5.77	91.11 ± 3.12	94.72 ± 2.36	0.40
angle 40		56.95 ± 11.34	83.22 ± 6.14	91.18 ± 3.21	94.59 ± 1.99	0.13
angle 0	$\frac{1}{3}$ N	53.61 ± 9.26	**80.01 ± 5.58**	**90.44 ± 2.36**	**94.52 ± 2.08**	0.53
angle 5		50.94 ± 9.08	76.54 ± 5.77	88.84 ± 3.88	93.58 ± 3.03	0.53
angle 10		**53.74 ± 4.92**	78.48 ± 5.29	88.97 ± 2.36	93.65 ± 2.36	0.94
angle 20		53.07 ± 9.26	78.14 ± 6.33	88.90 ± 4.54	93.72 ± 3.21	0.27
angle 40		50.40 ± 7.37	78.34 ± 5.67	89.51 ± 4.25	94.18 ± 2.93	0.40
angle 0	$\frac{1}{4}$ N	**48.86 ± 8.60**	75.07 ± 6.90	87.63 ± 3.69	92.51 ± 1.70	1.20
angle 5		48.46 ± 7.09	73.66 ± 5.86	87.50 ± 3.50	92.25 ± 2.84	1.34
angle 10		45.99 ± 6.62	**75.20 ± 6.71**	**88.37 ± 3.59**	**92.91 ± 1.89**	1.74
angle 20		45.79 ± 6.33	74.06 ± 8.51	86.50 ± 5.29	91.91 ± 2.74	0.80
angle 40		47.86 ± 10.40	**75.20 ± 7.85**	87.50 ± 3.88	92.65 ± 2.65	0.53
angle 0	$\frac{1}{5}$ N	**46.66 ± 7.56**	**74.20 ± 7.37**	**86.76 ± 5.29**	**92.31 ± 3.31**	0.80
angle 5		45.19 ± 7.18	73.86 ± 7.47	86.10 ± 5.10	91.64 ± 2.93	0.67
angle 10		41.11 ± 9.93	70.25 ± 8.79	83.96 ± 6.43	90.71 ± 3.31	0.40
angle 20		41.51 ± 10.68	71.66 ± 10.78	84.49 ± 6.05	90.44 ± 4.25	0.80
angle 40		43.85 ± 11.91	72.13 ± 12.38	84.56 ± 7.85	91.64 ± 4.25	1.20
angle 0	$\frac{1}{6}$ N	41.24 ± 6.71	69.85 ± 5.58	**84.56 ± 4.82**	90.24 ± 3.40	1.60
angle 5		**43.98 ± 5.67**	**70.66 ± 7.28**	83.29 ± 7.00	90.57 ± 2.55	0.94
angle 10		39.30 ± 7.37	69.39 ± 6.81	83.89 ± 5.96	**90.91 ± 3.59**	1.07
angle 20		40.51 ± 6.43	70.45 ± 7.56	84.02 ± 5.01	90.57 ± 2.74	1.34
angle 40		41.04 ± 9.83	67.78 ± 9.26	83.36 ± 7.09	90.57 ± 3.88	2.14
angle 0	$\frac{1}{7}$ N	34.36 ± 9.45	64.17 ± 10.02	80.01 ± 6.90	88.17 ± 4.63	2.54
angle 5		**35.76 ± 7.66**	**65.31 ± 6.90**	**80.68 ± 3.88**	**88.44 ± 2.93**	1.60
angle 10		35.29 ± 9.26	63.10 ± 7.00	79.88 ± 6.52	88.24 ± 4.54	2.14
angle 20		33.29 ± 6.81	62.03 ± 9.08	79.01 ± 5.48	**88.44 ± 4.25**	1.74
angle 40		35.16 ± 8.32	62.50 ± 8.41	80.41 ± 4.25	87.57 ± 4.54	2.67

Appendix 3. Results with modified validation data amount



(a) 1.0 mm

(b) 1.5 mm

(c) 2.0 mm

(d) 2.5 mm

Figure 23: KNEEL / optimized KNEEL / KNEEL+ trained with and without pre-training.

Appendix 4. Full comparison between KNEEL, optimized KNEEL, KNEEL+ with new validation data amount

method	# labeled in train set	1 mm	1.5 mm	2 mm	2.5 mm	% out
KNEEL optimized	N	58.16 ± 8.51	82.49 ± 4.92	92.25 ± 2.84	96.52 ± 1.32	0.13
KNEEL optimized w/o lc	N	55.61 ± 10.78	81.48 ± 7.47	91.44 ± 2.65	95.32 ± 1.70	0.13
KNEEL+ (val fixed)	N	**66.64 ± 10.30**	87.83 ± 5.10	93.85 ± 1.89	96.66 ± 1.13	0.00
KNEEL+ 750 unlabeled	N	66.11 ± 7.66	**88.37 ± 3.03**	**94.32 ± 1.80**	**97.06 ± 1.51**	0.27
KNEE lc	N	45.25 ± 8.79	72.59 ± 6.24	85.76 ± 3.88	90.64 ± 1.89	0.80
KNEE w/o lc	N	34.56 ± 2.74	63.84 ± 2.93	81.55 ± 1.51	89.10 ± 0.47	1.20
KNEEL optimized	$\frac{1}{2}$ N	50.67 ± 12.07	77.60 ± 7.92	87.87 ± 4.71	92.67 ± 1.70	0.53
KNEEL optimized w/o lc	$\frac{1}{2}$ N	45.07 ± 12.82	72.27 ± 10.56	86.93 ± 4.90	92.27 ± 1.89	0.53
KNEEL+ (val fixed)	$\frac{1}{2}$ N	59.87 ± 16.78	84.00 ± 5.66	**93.47 ± 3.58**	**96.93 ± 1.70**	0.53
KNEEL+ 750 unlabeled	$\frac{1}{2}$ N	**62.40 ± 15.46**	**84.80 ± 6.03**	92.67 ± 2.07	96.13 ± 1.32	0.27
KNEE lc	$\frac{1}{2}$ N	33.42 ± 5.10	61.76 ± 4.35	78.41 ± 2.93	86.16 ± 1.42	2.27
KNEE w/o lc	$\frac{1}{2}$ N	19.72 ± 3.50	46.66 ± 3.59	65.78 ± 2.65	79.14 ± 1.51	3.07
KNEEL optimized	$\frac{1}{3}$ N	50.00 ± 7.35	76.20 ± 4.24	86.20 ± 1.98	92.00 ± 2.26	0.80
KNEEL optimized w/o lc	$\frac{1}{3}$ N	40.60 ± 8.77	68.40 ± 9.05	84.20 ± 6.51	90.60 ± 3.68	0.40
KNEEL+ (val fixed)	$\frac{1}{3}$ N	**59.80 ± 11.03**	82.60 ± 5.94	**93.00 ± 3.11**	94.80 ± 2.26	0.40
KNEEL+ 750 unlabeled	$\frac{1}{3}$ N	59.00 ± 15.56	**84.80 ± 5.66**	92.40 ± 3.96	**95.20 ± 2.26**	0.40
KNEE lc	$\frac{1}{3}$ N	24.47 ± 3.03	51.47 ± 4.73	69.85 ± 2.55	80.88 ± 2.08	3.74
KNEE w/o lc	$\frac{1}{3}$ N	10.29 ± 1.51	29.08 ± 0.66	49.73 ± 0.95	64.84 ± 1.13	4.41
KNEEL optimized	$\frac{1}{4}$ N	42.89 ± 7.82	71.32 ± 8.56	86.84 ± 5.95	92.63 ± 2.98	2.11
KNEEL optimized w/o lc	$\frac{1}{4}$ N	36.84 ± 5.21	64.21 ± 3.72	82.89 ± 2.61	88.95 ± 1.49	1.58
KNEEL+ (val fixed)	$\frac{1}{4}$ N	55.00 ± 9.30	81.32 ± 8.56	92.11 ± 2.98	94.47 ± 1.12	1.05
KNEEL+ 750 unlabeled	$\frac{1}{4}$ N	**55.79 ± 8.93**	**85.26 ± 8.19**	**92.89 ± 5.58**	**95.79 ± 2.23**	0.53
KNEE lc	$\frac{1}{4}$ N	19.79 ± 1.13	43.72 ± 0.95	63.37 ± 1.13	75.67 ± 0.38	5.48
KNEE w/o lc	$\frac{1}{4}$ N	8.69 ± 2.46	25.33 ± 1.80	44.92 ± 3.03	59.29 ± 1.42	7.89
KNEEL optimized	$\frac{1}{5}$ N	33.33 ± 4.71	64.00 ± 11.31	81.67 ± 8.01	89.67 ± 3.30	1.33
KNEEL optimized w/o lc	$\frac{1}{5}$ N	28.33 ± 10.84	58.67 ± 5.66	77.33 ± 0.94	88.33 ± 3.30	2.67
KNEEL+ (val fixed)	$\frac{1}{5}$ N	46.00 ± 10.37	**78.67 ± 5.66**	**90.00 ± 2.83**	**94.00 ± 1.89**	1.33
KNEEL+ 750 unlabeled	$\frac{1}{5}$ N	**47.33 ± 14.14**	75.67 ± 9.90	87.67 ± 6.13	93.00 ± 2.36	1.33
KNEE lc	$\frac{1}{5}$ N	11.50 ± 0.76	31.95 ± 1.13	52.01 ± 2.65	67.45 ± 1.99	7.35
KNEE w/o lc	$\frac{1}{5}$ N	5.28 ± 1.04	17.11 ± 1.13	34.29 ± 1.04	51.00 ± 2.93	7.89
KNEEL optimized	$\frac{1}{6}$ N	32.80 ± 3.39	63.20 ± 0.00	82.00 ± 5.09	87.20 ± 1.13	3.20
KNEEL optimized w/o lc	$\frac{1}{6}$ N	24.80 ± 4.53	51.60 ± 3.96	70.80 ± 1.70	82.80 ± 1.70	3.20
KNEEL+ (val fixed)	$\frac{1}{6}$ N	**49.20 ± 13.01**	**77.60 ± 6.79**	**88.00 ± 4.53**	91.20 ± 2.26	0.80
KNEEL+ 750 unlabeled	$\frac{1}{6}$ N	47.60 ± 10.75	76.80 ± 9.05	**88.00 ± 4.53**	**91.60 ± 1.70**	0.80
KNEE lc	$\frac{1}{6}$ N	8.82 ± 0.57	26.40 ± 0.28	43.32 ± 0.38	58.62 ± 0.47	9.09
KNEE w/o lc	$\frac{1}{6}$ N	3.34 ± 1.51	10.76 ± 2.93	25.07 ± 2.74	40.64 ± 3.21	9.63
KNEEL optimized	$\frac{1}{7}$ N	26.36 ± 7.71	57.73 ± 9.64	78.64 ± 1.93	86.82 ± 0.64	5.45
KNEEL optimized w/o lc	$\frac{1}{7}$ N	21.36 ± 0.64	44.55 ± 5.14	62.27 ± 3.21	75.45 ± 1.29	3.64
KNEEL+ (val fixed)	$\frac{1}{7}$ N	**45.45 ± 12.86**	**76.36 ± 3.86**	**89.09 ± 2.57**	**93.18 ± 0.64**	0.91
KNEEL+ 750 unlabeled	$\frac{1}{7}$ N	42.73 ± 12.86	75.91 ± 9.64	87.73 ± 1.93	91.36 ± 0.64	2.73
KNEE lc	$\frac{1}{7}$ N	5.28 ± 2.36	16.58 ± 5.48	31.62 ± 8.22	46.19 ± 7.85	11.50
KNEE w/o lc	$\frac{1}{7}$ N	1.40 ± 0.66	4.81 ± 2.65	11.36 ± 4.73	20.86 ± 5.29	22.33