A Web-Based Software for Training and Quality Assessment in the Image Analysis Workflow for Cardiac T1 Mapping MRI

Edvarda Eriksen*†, Steven Hicks‡§, Michael Alexander Riegler‡,‖ Pål Halvorsen‡§, and Valentina Carapella¶
*Simula Research Laboratory, Norway †University of Oslo, Norway ‡SimulaMet, Norway
§Oslo Metropolitan University, Norway ¶King's College London, United Kingdom
‖Kristiania University College, Norway
Contact email: edvarda.eriksen@outlook.com

***Abstract*—Medical practice makes significant use of imaging scans such as Ultrasound or Magnetic Resonance Imaging (MRI) as a diagnostic tool. They are used in the visual inspection or quantification of medical parameters computed from the images in post-processing. However, the value of such parameters depends much on the user's variability, device, and algorithmic differences. In this paper, we focus on quantifying the variability due to the human factor, which can be primarily addressed by the structured training of a human operator. We focus on a specific emerging cardiovascular MRI methodology, the T1 mapping, that has proven useful to identify a range of pathological alterations of the myocardial tissue structure. Training, especially in emerging techniques, is typically not standardized, varying dramatically across medical centers and research teams. Additionally, training assessment is mostly based on qualitative approaches. Our work aims to provide a software tool combining traditional clinical metrics and convolutional neural networks to aid the training process by gathering contours from multiple trainees, quantifying discrepancy from local gold standard or standardized guidelines, classifying trainees output based on critical parameters that affect contours variability.**

***Index Terms*—Cardiovascular MRI, T1 mapping MRI, quality assessment, deep learning, image analysis training, standardisation.**

I. Introduction

In recent years, machine learning-based algorithms have been used to perform a variety of different tasks within medicine [1]. Deep learning, for example, has already shown much success in aiding medical doctors to perform diagnosis on diseases including different types of cancers and other conditions [2], [3]. For the most part, machine learning algorithms are only as good as the data used to train them. Therefore, quality datasets are essential as they directly affect the performance of a given model. In the medical domain, datasets are usually created by medical doctors who collect and annotate each element in a given set. Due to the time-consuming process of annotating large datasets, and a general lack of medical doctors, it is often tough to get high-quality datasets [4]. Therefore, making this process more efficient could significantly improve the availability of quality medical datasets. This would not only aid in making better machine-learning models but could also be used in the medical community to train students in performing different tasks such as segmentation. Using automatic methods to support the training of medical experts is a somewhat neglected topic. There does not exist much work on how machine learning can be used to make the training of junior doctors better and more efficient.

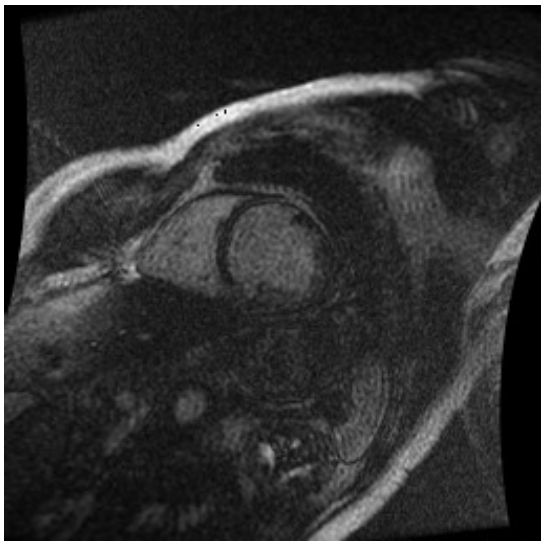

Fig. 1: Example MRI taken from the Sunnybrook Cardiac Data (SCD) [5] dataset. Note that the presented image has been resized for the purpose of this figure.

In this work, we propose a system that targets a specific area within this problem, specifically, creating annotations. As a use case, we chose T1 mapping in cardiac Magnetic Resonance Imagings (MRIs), which requires a precise segmentation of the left ventricle of the heart.

T1 mapping in cardiac MRIs is a rapidly emerging technique in the clinical setting to identify microstructural defects associated with a range of cardiovascular disease [6]. It achieves this by quantifying the spin-lattice relaxation time (T1) of protons in the water molecules of the biological tissue. This is measured by hitting the protons with a radio frequency signal and seeing how long it takes (in milliseconds) for the protons to return to their original state during an MRI procedure. T1 times are determined by the proportion of water content in tissue, in addition to its compartmentalization, so that each tissue type (such as blood or fat) show a range of different T1 values. These ranges, however, change drastically at different magnetic field strengths and may also depend on the specific MRI sequence used to measure them [7]. T1 mapping is the process of mapping T1 times to the individual pixels for a given MRI. These mappings are used to visualize the MRI image to easily distinguish

between the different types of tissue, including blood, fat, and muscle. Standardization of image acquisition, post-processing, and interpretation is crucial to ensure the consistency and reproducibility of any medical image analysis process, as well as guaranteeing a common ground for clinical assessment. Clinical use of cardiovascular T1 mapping is no exception, and a standardization task force has already provided the initial guidelines [8].

A T1 mapping of cardiac MRIs is a process which requires a segmentation of the inner and outer left ventricle of the heart (Figure 1 shows an example MRI image of the left ventricle). These segmentations are used to calculate the wall thickness (WT) and the T1 value, which both set the basis for detecting a range of abnormalities. Placing these segmentations requires trained observers who understand how to interpret and analyze a cardiac MRI. Observers are commonly trained through a training program consisting of segmentation tasks on an expert segmented training dataset, which is made up of segmentations created by a consensus of multiple expert observers. However, labeled datasets are limited, and evaluating each student's segmentation is time-consuming and lacks the immediate feedback needed for rapid improvement. Therefore, we propose a system that aids in the training of new observers. For the system to be truly useful, it needs to incorporate the following requirements:

1) Produce a standardized report of the training progress and completion for accreditation for an individual trainee, or team of trainees.
2) Include a centralized approach to compare the performance of a single trainee, as well as a team of students, against the accepted reference for a training dataset.
3) Automatically generate an average contour either from expert analysis (typically called the *consensus*) or the average of team results.

With these requirements in mind, we present a software tool that aids the training of new observers by automatically evaluating their work by comparing against an automatically generated "expert" consensus using deep convolutional neural networks (CNNs).

The novelty of our software is two-fold. Firstly, it focuses on the evaluation of trainees working on T1 mappings in cardiac MRIs, not just from an individual point of view, but also from the viewpoint of multiple operators (who typically work in a clinical research unit). This combination of individual and team evaluation stems from the fundamental need for consistency in the performed analysis, not just concerning general guidelines, but also within a clinical unit.

Secondly, the expert observer segmentations used to evaluate a trainees (or group) segmentation is automatically generated by a deep CNN. This CNN is modeled and trained to imitate the knowledge of several expert observers. By using pre-existing training datasets, we teach the model to estimate an outline of the outer (epicardium) and inner (endocardium) wall of the myocardium for a given MRI together with a trainees contour. The presented approach allows for instant feedback to trainees without having an expert annotate new MRI images since the CNN back-end can get the expert feedback from based on the student annotation. Furthermore, another advantage is that a training facility would not have to rely on expert contoured MRI datasets.

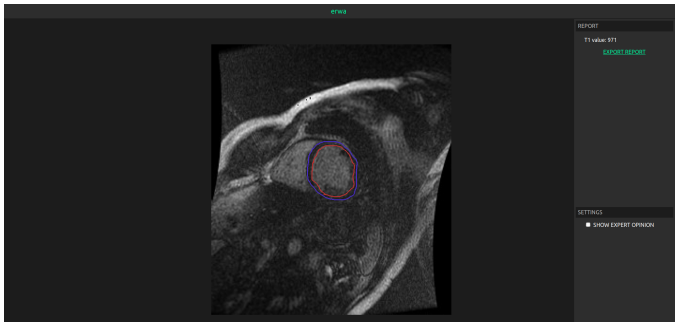

Fig. 2: Screenshot of the Trainee's Dashboard after the T1 analysis is uploaded.

In the process of analyzing T1 mappings in cardiac MRIs, two measurements are typically recorded; The average myocardial T1 value measured in milliseconds, and the average myocardial WT measured in pixels or millimeters. The value of T1 is the most clinically relevant in these types of scans, whereas WT is mostly employed as a quality assessment metric; its value is loosely correlated with the average T1. Regarding the underlying research carried out alongside the software development, we present a combination of traditional metrics of evaluation (such as average myocardial T1 and WT) that are easily understandable to a clinical audience.

The rest of this paper is structured as follows. In Section II, we give a brief look at some previous works done in this field. This section mostly focuses on work done for the automatic segmentation of the left ventricular myocardial ring. Section III further details how the developed software may support the cardiovascular magnetic resonance imaging (CMR) community as a whole, and the main contributions we aim to deliver with this research. Section IV describes our software, showing how it may be used for the training of new observers. Then, in Section V, we look at the architecture and implementation details of this software and discuss in detail how it works. Section VI shows the experiments performed for the underlying *expert* observation CNN model used to evaluated trainee segmentations. Lastly, Section VII concludes this paper with a summary of our findings and a brief discussion on future work.

II. Related Works

Our software aims to aid students in training to become expert observers of T1 mappings in cardiac MRIs. It does this by providing immediate feedback on the students' performance and providing a progress report for individuals and classes of students. Image post-processing of T1 mapping in MRIs requires the delineation of a region of interest (ROI), that is, placing contours. The most common case is that of the left ventricular myocardial ring. Depending on the software, contours can be set manually or semi-automatically. However, manual inspection should be carried out to ensure that the

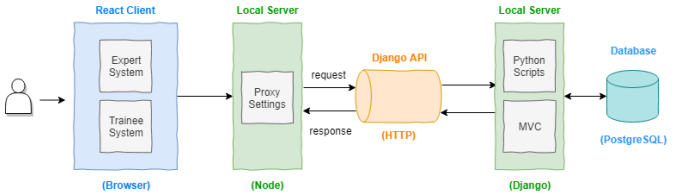

Fig. 3: A diagram showing how data is transferred across the system, starting with the user and ending with the Django based server. From the left, we see that a user uploads a segmentation together with the associated training image. This is then passed through the local Node based proxy server to the centralized Django server. The image together with the contour is stored in the database and evaluated by the CNN "expert". The evaluation is then passed back to the user where he/she is presented with their performance metrics.

contours do not include regions outside of the myocardium (such as the outer pericardial fat or inner blood pool). Some more thorough quality assessment also label regions in the myocardium affected by imaging artifacts to exclude them from the analysis.

Over the years, multiple approaches for automatically placing these contours have been proposed. In 2014, Hu et al. [9] proposed a method for automatically segmenting the left ventricular myocardial using local binary fitting models and dynamic programming techniques. Overall, their approach shows good results, yet they struggle to segment the overlap between intensity distributions within the cardiac regions. Abdelfadeel et al. [10] use maximally stable extremal regions to segment the left ventricle of cardiac MRIs. Their achieves a DICE metric of 0.88 on the Medical Image Computing and Computer Assisted Intervention (MICCAI) 2009 challenge database [5]. Similar to our method, Zreik et al. [11] propose a technique based on deep CNNs, where they try to segment the left ventricle in cardiac CT images. Their method uses a combination of three different CNNs, each detecting the presence of the left ventricle in the axial, coronal, and sagittal independently. Our work, however, differs from these approaches as we try to measure the quality of a student contour by generating an expert consensus based on the student contour and MRI image using a deep CNN.

III. The Research Contribution

The development of the software tool started from the specific requirements gathered from a team of researchers working in the field. Thus, first of all, it is a tool for research support. In this respect, by allowing more efficient and consistent monitoring of quality assessment in cardiac MRI image analysis, it has a fundamental impact on the quality of the research produced at that center. However, this has allowed us to carry out novel research in the field of quality assessment using deep learning methods, specifically to automate the generation of contours based on expert observer analysis (in other words, generating an expert consensus). This approach to quality assessment based on the comparison of contours, rather than point measures such as T1 or WT, has not yet been fully explored in the CMR community, with some exceptions [12]. We believe that this aspect of our work will contribute to a more quantitative, detailed, and standardized approach to quality assessment of image analysis done on T1 mapping in cardiac MRIs.

IV. System Description and Usage

We have divided the application into two user groups: (a) experts/supervisors and (b) trainees. The system expects both users to upload files of the file types Digital Imaging and Communications in Medicine (DICOM) [13], that contain the original T1 mapping MRI data, and Interactive Data Language (IDL) SAV Files [14], that includes the image segmentation, that is the black and white binary mask representing the ROI enclosed by the contours (specifically, the image segmentation). The system is implemented as a web-based application, comprising of a back-end server and front-end web interface (which will be further discussed in Section V).

A. The Supervisor System

The supervisor (expert) can upload a pre-generated consensus of experts into the system, which will give the trainees the ability to compare their analysis with that of the gold-standard. Also, the supervisor can enable an automatically generated consensus to be an evaluation option for the trainees. This allows for the use of non-contoured datasets, which increases the overall training material. The system will also allow supervisors to view the overall progress of a class of trainees by getting an automatically generated report about their collective progress. The report is produced in a PDF format and provides various graphs and summary metrics about individual trainees and the class as a whole.

B. The Trainee System

As previously stated, the main purpose of this software is to train observers in contouring the endocardium and epicardium of the left ventricle and giving immediate feedback by comparing against an automatically generated "expert" consensus produced by a deep CNN. For an example use case scenario, we imagine a trainee who has been working on contouring an MRI as part of their training dataset. The trainee uploads the T1 mapping of the MRI (or DICOM file) along with their produced segmentation into the system. The system will then start extracting key values from the DICOM [13] file, of which the trainee will be presented with a visual representation of their work, as well as the T1-value and the WT. Then, to get an understanding of their progress, the trainee may compare against the systems generated "expert" contour consensus by getting a visual representation of the overlay between the two contours (trainee produced contour and CNN produced contour), showcasing the regions of the discrepancy between the two. Furthermore, the user will get some key metrics regarding how well their contour compares to the one generated by the system. These evaluation metrics will be used to clearly state how well the trainees' segmentation compares to that of the consensus.

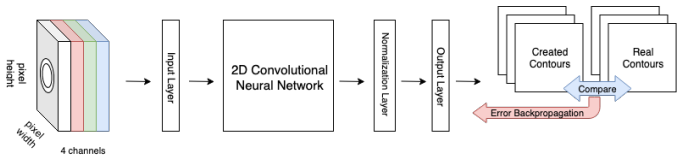

Fig. 4: A diagram showing the entire process for training the "expert" consensus CNN used to evaluated trainees. Starting from the left, we see that the input image (comprised of the student contour and MRI) is resized to $224 \times 224 \times 4$ before being put into the model. Then, the model outputs a vector with size $52,176$, which is then reshaped into the same shape as the input image.

The trainee can also generate a basic T1 training report to track the progress of their repeated analysis over the same dataset. This is produced in a PDF format and provides an overview of the progress with graphs and summary metrics showing the discrepancies between their analysis and the consensus. This report may also be used to track the development of a team of trainees, showing how a class improves over time.

V. Software Architecture and Implementation

Due to the sensitive nature of the data used for training, as well as the geographical localization of our target demographic, the system is only meant to be run over a local area network (LAN). That is, it is not meant to be available outside the area of deployment. With this in mind, our web application is managed by two locally run servers. One is based on Node.js [15] acting as an intermediate proxy which is responsible for handling the front-end logic, and the other is based on the python web-framework Django [16] which interacts with the database and deep learning model. A screenshot of the web-based graphical user interface (GUI) is shown in Figure 2, and the overall architecture is shown in Figure 3.

The web-based client is built using React [17] (a popular JavaScript library used to build front-end interfaces), which communicates directly with the Node.js based server. This server acts as a proxy that redirects all HTTP requests and responses to the Django server. From here, Django may retrieve images from the database or evaluate uploaded contours against the contours produced by the CNN model or expert consensus. Furthermore, the Django-based server is also responsible for calculating the evaluation metrics used to assess a trainee's performance on contouring the left ventricle. The underlying CNN is implemented using the deep learning framework Keras [18] with a TensorFlow [19] back-end. The CNNs purpose is to automatically infer the endocardium and epicardium contours based on a given MRI and student contour.

VI. Automated Feedback Using CNNs

Part of what makes the system novel is the automatic generation of an "expert" contour consensus using deep CNNs. The purpose of automating this process is to fulfill the need for quick and consistent feedback to the trainees using the system, in addition to not needing a fully expert contoured dataset. This section will cover the various experiments done in producing this "expert" observer CNN, and qualitative analysis on the models produced contours. But first, we will give a short description of the dataset used to train our models.

A. Dataset Details and Preprocessing

The dataset consists of 42 fully anonymized single mid-ventricular short-axis native T1-maps, half of which come from healthy volunteers, while the other half from patients with an acute myocardial injury. Each T1-map comes as a DICOM file (a standard for medical imaging data). A typical DICOM file contains a wide variety of different information ranging from simple meta-data (such as information about the MRI sequence or details about the patient), to full MRI images. Because we are only interested in the MRI data, the MRI images had to be extracted before training the neural network. This was done using the python library called pydicom [20], which is a popular library for working with DICOM files. The extracted images vary in size, ranging from 384×264 to 384×344 pixels.

For the ground truth contours, the dataset includes several contours per DICOM file. Each MRI has been contoured by several experts to form a consensus between them. Overall, there are two sets of contours; one for the inner myocardial wall (endocardium), and one for the outer (epicardium). These contours come in a special file written in a proprietary programming language called IDL. The contents of these files are coordinates that correspond to the contours of the related MRI image. For training purposes, we converted these files to a pixel activated vector which was used as the ground truth for our CNN experiments.

B. CNN Architecture and Training Configuration

The network used to generate segmentations is a modified version of the VGG-16 [21] based model implemented in Keras. To decide on which architecture to use, we experimented with numerous of the Keras-based implementations of popular CNN architectures (such as, ResNet [22], Inception [23], DensNet [24]), but we ended up using VGG-16 as it showed the best performance. For input, the model expects an image consisting of four channels. The first three channels consist of the R (red), G (green), and B (blue) color channels of the extracted MRI image (extracted as previously described). The student contour represents the fourth image channel. The input image is then resized $224 \times 224 \times 4$ before being passed into the model, which outputs a vector with the size $50,176$ (224×224) which represents the segmentation of the left ventricular myocardial ring. Each value in the vector represents one pixel of the output segmentation, and each pixel can be either "on" or "off", meaning it consists of $50,176$ binary values being either 1 or 0. To get the final segmentation, we chop the output vector into 224 pieces and stack them on top of each other before turning "on" pixels white and "off" pixels black.

(a) CNN system (b) Expert observer

Fig. 5: The segmentation results produced by the system's CNN and an expert observer.

For training, we used the Keras based implementation of VGG-16 and trained it from scratch. We replace the classification block of the original model with a custom block consisting of one 2D global average layer, a normalization layer that squashes the input values between 0 to 1, then a final fully-connected layer consisting of $50,176$ nodes (one for each pixel). The model was trained using mean absolute error (MAE) to calculate loss and Nadam [25] to optimize the weights with a learning rate of 0.004, 0.900 for β_1, and 0.999 for β_2. The ground truth is created from expert segmentations made by multiple observers. This includes several expert segmentations for a single MRI image, which are all used in the training process to let the network average an "expert" consensus. To ensure robust results, each experiment was run using three-fold cross-validation. A diagram explaining the entire training process can be seen in Figure 4.

Figure 5 shows some examples of segmentations produced by the underlying CNN model compared to that of the actual expert-created segmentation. We see that the CNN-created segmentation is quite similar to that of the expert but still requires some improvements before being deployed into a fully functional system.

VII. Conclusion and Future Work

In this paper, we presented a software tool meant to aid students in contouring the inner and outer wall of the left ventricle of the heart. The system automatically evaluates the trainees' segmentation by comparing it against an "expert" consensus generated by a deep CNN. This allows for immediate feedback on their submitted segmentation and removes the need for a fully contoured expert dataset. Furthermore, the system presents an overview of the student progress using a variety of different metrics and graphs. The underlying CNN used for generating the "expert" segmentation is based on a VGG-16 architecture and trained on a dataset consisting of segmentations made by multiple cardiologists.

For future work, we aim to support more options for generating an expert consensus, in addition to improving the existing method. Furthermore, we intend to get this software into the hands of real trainees and observers to gain further feedback and real-world evaluation of this tool.

References

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, p. 44, 2019.

[2] "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[3] A. Esteva, B. Kuprel, R. A. Novoa *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] M. Riegler, M. Lux, C. Griwodz *et al.*, "Multimedia and medicine: Teammates for better disease detection and survival," in *Proceedings of the 24th ACMMM*, 2016, pp. 968–977.

[5] P. Radau, Y. Lu, K. Connelly *et al.*, "Evaluation framework for algorithms segmenting short axis cardiac mri," July 2009.

[6] D. R. Messroghli, J. C. Moon, V. M. Ferreira *et al.*, "Clinical recommendations for cardiovascular magnetic resonance mapping of t1, t2, t2* and extracellular volume: A consensus statement by the society for cardiovascular magnetic resonance (scmr) endorsed by the european association for cardiovascular imaging (eacvi)," *Journal of Cardiovascular Magnetic Resonance*, vol. 19, no. 1, p. 75, Oct 2017.

[7] S. K. Piechnik, V. M. Ferreira, E. Dall'Armellina *et al.*, "Shortened modified look-locker inversion recovery (shmolli) for clinical myocardial t1-mapping at 1.5 and 3 t within a 9 heartbeat breathhold," *Journal of Cardiovascular Magnetic Resonance*, vol. 12, no. 1, p. 69, Nov 2010.

[8] J. C. Moon, D. R. Messroghli, P. Kellman *et al.*, "Myocardial t1 mapping and extracellular volume quantification: a society for cardiovascular magnetic resonance (scmr) and cmr working group of the european society of cardiology consensus statement," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, p. 92, Oct 2013.

[9] H. Hu, Z. Gao, L. Liu *et al.*, "Automatic segmentation of the left ventricle in cardiac mri using local binary fitting model and dynamic programming techniques," *PLOS ONE*, vol. 9, no. 12, pp. 1–17, Dec 2014.

[10] M. A. Abdelfadeel, S. ElShehaby, and M. S. Abougabal, "Automatic segmentation of left ventricle in cardiac mri using maximally stable extremal regions," in *Proceedings of the 7th CIBEC*, Dec 2014, pp. 145–148.

[11] M. Zreik, T. Leiner, B. D. de Vos *et al.*, "Automatic segmentation of the left ventricle in cardiac ct angiography using convolutional neural networks," in *Proceedings of the 13th ISBI*, April 2016, pp. 40–43.

[12] A. Suinesiaputra, D. A. Bluemke *et al.*, "Quantification of lv function and mass by cardiovascular magnetic resonance: multi-center variability and consensus contours." *Journal of cardiovascular magnetic resonance*, vol. 17, p. 63, 2015.

[13] National Electrical Manufacturers Association. (2019) Dicom standard. Available: https://www.dicomstandard.org/

[14] Harris Geospatial Solutions Inc. (2019) The save procedure. Available: https://www.harrisgeospatial.com/docs/SAVE.html

[15] Node.js Foundation, "Node.js," 2019. Available: https://nodejs.org/en/

[16] Django Software Foundation. (2019) Django - the web framework for perfectionists with deadlines. Available: https://www.djangoproject.com/

[17] Facebook Inc. (2019) React - a javascript library for building user interfaces. Available: https://reactjs.org/

[18] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[19] M. Abadi, P. Barham, J. Chen *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th OSDI*, 2016, pp. 265–283.

[20] D. Mason, scaramallion, rhaxton *et al.*, "pydicom/pydicom: 1.2.2," Jan. 2019.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.

[22] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proceedings of the 29th CVPR*, June 2016, pp. 770–778.

[23] C. Szegedy, P. Sermanet, S. Reed *et al.*, "Going deeper with convolutions," in *Proceedings of the 28th CVPR*, June 2015, pp. 1–9.

[24] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the 30th CVPR*, July 2017, pp. 2261–2269.

[25] T. Dozat, "Incorporating nesterov momentum into adam," 2015.