Lightweight U-Net for Lesion Segmentation in Ultrasound Images

Yingping Li, Emilie Chouzenoux, Benoit Charmettant, Baya Benatsou, Jean-Philippe Lamarque, Nathalie Lassau

LIGHTWEIGHT U-NET FOR LESION SEGMENTATION IN ULTRASOUND IMAGES

Yingping Li[2,3] *Emilie Chouzenoux*[3] *Benoit Charmettant*[2] *Baya Benatsou*[1,2]
Jean-Philippe Lamarque[1,2] *Nathalie Lassau*[1,2]

[1] Imaging Department Gustave Roussy, Université Paris Saclay, 94805 Villejuif, France
[2] Biomaps, UMR1281 INSERM, CEA, CNRS, Université Paris-Saclay, 94805 Villejuif, France
[3] OPIS, Inria Saclay, Université Paris-Saclay, 91190 Gif-sur-Yvette, France

ABSTRACT

Acquiring ultrasound images of suspected lesion areas allows radiologists to monitor the cancer development of patients. The goal of this paper is to provide an automatic lesion segmentation tool for assisting them on the analysis of ultrasound images, by relying on recent neural network methods. Specifically, we perform a comparative study for the segmentation of 348 ultrasound image pairs acquired in 19 centers across France, displaying different tumor types. We show that, with a careful hyperparameter tuning, U-net outperforms other state-of-the-art networks, reaching a Dice coefficient of 0.929. We then propose to introduce group convolution into U-net architecture. This leads to a lightweight network named Lighter U-net @128 that achieves comparable segmentation performance with obviously reduced model size, hence paving the way for an embedded integration within hospital environment. We made our code publicly available[1], for reproducibility purpose.

***Index Terms*—** Ultrasound images, lesion segmentation, U-net, lightweight network.

1. INTRODUCTION

Acquiring ultrasound images for tracking lesions is a non-invasive imaging routine tool for radiologists to monitor the tumor development during the cancer treatment process. Segmenting the lesion boundaries from the acquired ultrasound images is the first necessary step to perform radiomic-based study, that is to investigate new prognostic biomarkers inside the segmented tumor, and thus appears as an important step for further diagnosis and treatment [1]. Manual lesion segmentation process is often very tedious, time-consuming and different segmentation qualities may come out from different radiologists. Hence automatic image segmentation methods are necessary.

Since Jonathan *et al.* [2] proposed the fully convolutional networks (FCN) which extended a classification network to pixel-level segmentation in an end-to-end manner, many neural network methods have been proposed for image segmentation [2–5]. They even become the primary option for tackling medical image segmentation tasks [6]. Paper [7] reviewed the literature and grouped the ultrasound image segmentation neural networks into six classes depending on the retained network architectures or training methods: FCN [2, 8], encoder-decoder networks (e.g., U-net, DeepLab v3+) [3, 4], recurrent neural networks [9], generative adversarial networks [10], weakly supervised learning [11] and deep reinforcement learning methods [12]. Among these 6 classes, encoder-decoder networks (48%) and FCN (24%) appear as the most common used architectures in papers published in recent years [7]. Let us emphasize that encoder-decoder networks with skip connections have also shown their superiority in several other applications of medical image segmentation [13].

In this work, we perform a comparative analysis of state-of-the-art neural network based segmentation techniques, namely two encoder-decoder networks (U-net [3], DeepLab v3+ [4]), and FCN [2], for performing the segmentation of lesions in ultrasound images arising from a multicentric and multipathology study [14,15] supervised by the radiology unit of Gustave Roussy Institute, one of the leading cancer center and research institute in the world. We discuss the process of hyperparameter tuning for the best performing network, U-net. We then propose to reduce the model size with the aim to favor its embedding in medical systems without sacrificing the segmentation performance. In particular, we show that our lightweight version of U-net named Lighter U-net @128, achieves comparable segmentation performance on our data but more than two times less model parameters, making it easier to be integrated within the hospital computing tools.

The paper is organized as follows: Section 2 describes the dataset, Section 3 presents our methodology for the comparative analysis, and our approach for lightening the network memory occupation. Section 4 presents the results and discusses the setting for the hyperparameters. Finally, Section 5 concludes the paper.

[1] `https://github.com/Yingping-LI/Light-U-net`

2. DATASET DESCRIPTION

The data used in this paper follows the standardization of dynamic contrast-enhanced ultrasound (DCE-US) which is used for predicting outcomes of antiangiogenic therapy for solid tumors. More details can be found in [14, 15]. The original study included 539 patients from 19 centers across France (11 comprehensive cancer centers and 8 teaching hospitals) between October 2007 and March 2010. The involved patients suffered from different types of tumors, such as metastatic breast cancer, melanoma, colon cancer, gastrointestinal stromal tumors, renal cell carcinoma and primary hepatocellular carcinoma tumors. They were all treated with antiangiogenic therapy, and the DCE-US examinations were performed at baseline (day 0) as well as on days 7, 15, 30, 60. Such examination involved taking ultrasound images, recording the perfusion curve during 3 minutes after injection of a contrast agent and so on. For each patient, the examination only focused on one target tumor in the body.

We only have access to part of the complete dataset (348 patients from the original 539 patients). For each patient, we focus on the two ultrasound images in B-mode in this retrospective dataset, which were taken in 2 mutual-orthogonal directions (front and side) at baseline (day=0) with the tumors measured by electronic calipers in light-blue by the radiologists, see Figure 2. The images are used as color images (RGB, 3 channel) in our study. Besides, we also need annotation of the lesion in each image to train the neural networks. The annotation task has been performed manually by a PhD student, then validated by an expert with 28 years experiences in acquiring and analysing ultrasound images. We randomly split the dataset into 208, 70, and 70 image pairs as training, validation, and test data. The two images in each image pair are fed into the network separately, i.e. the network takes a single image as input. The training and validation datasets are used for training and tuning the network. The test dataset, never seen by the network during training/validation process, is used to evaluate fairly the network performance.

3. METHODOLOGY

3.1. State-of-the-art segmentation networks

We conduct a comparative analysis between several classical network architectures using their publicly available Pytorch implementation, namely FCN[2] [2] (first network to realize the end-to-end semantic segmentation), DeepLab v3+ [3] [4] (the newest version of a classic image segmentation network series), and U-net [4] [3] (state-of-the-art network for many medical image segmentation tasks). We consider the three versions of FCN (FCN-32s, FCN-16s, FCN 8-s). For DeepLab v3+, we choose to compare the most common used backbone ResNet-101 [16] and a lightweight backbone MobileNet [17].

The images are resized to $256 \times 256 \times 3$ pixels before being fed into the networks. Early stopping strategy (terminate training if the validation loss does not decrease for 15 consecutive epochs) is used to avoid overfitting, and batch normalization [18] is adopted to stabilize and accelerate the training process. The largest connected component of the predicted mask is then kept (as only one lesion is excepted in each image), and resized to recover the original height and width (380×646 or 432×646) as the final segmentation result. The hyperparameter tuning is discussed in detail in Section 4.

3.2. Proposed lightweight versions of U-net

In medical context, a lightweight network is always preferred for a reduced GPU cost and an easier embedding in hospital systems. We thus investigated the reduction of the network size while keeping similar segmentation performance, by introducing group convolution [19–21], itself generalizing depthwise separable convolution [17, 22]. We focus on the specific case of U-net, as it shows to reach the best segmentation performance in our comparative analysis.

Depthwise separable convolution consists of a depthwise convolution to extract the spatial correlations in each channel, followed by a pointwise convolution to extract cross-channel correlations. It relies on the assumption that the cross-channel correlations and spatial correlations can be mapped completely separately. As mentioned in [20], there are some discrete intermediate modules, called group convolutions nowadays [21], between regular convolution and depthwise separable convolution. In contrast with [20] which mentions to parameterize these intermediate modules by the number of independent channel-space segments (i.e., number of groups), we choose here to parametrize them instead by the number of channels (i.e., group size [21]) denoted $\mathcal{C}$, used for performing spatial convolutions at a time. Regular convolution (followed by a 1×1 convolution) corresponds to the extreme case when $\mathcal{C}$ equals to the total input channel number, which means performing spatial convolutions for all channels simultaneously. Depthwise separable convolution refers to the other extreme case when $\mathcal{C} = 1$, that is spatial convolution is performed for each input channel separately.

Setting group convolution with different values for group size $\mathcal{C}$ seems attractive, because it may be in accordance with the reality that only parts of the channels are related and should be performed for spatial convolutions simultaneously, thus reducing the model size and at the same time keeping the network capability. Based on this conjecture, we introduce group convolutions into U-net and propose some lightweight versions for it. The general network architecture is illustrated in Figure 1. The different versions can be obtained by specific choices of "Conv set 1" and "Conv set 2", as shown hereafter:
U-net: "Conv set 1" and "Conv set 2" correspond to regular

[2] https://github.com/pochih/FCN-pytorch

[3] https://github.com/jfzhang95/pytorch-deeplab-xception

[4] https://github.com/milesial/Pytorch-UNet

3×3 convolutions.

Light U-net: "Conv set 1" corresponds to regular 3×3 convolution while "Conv set 2" corresponds to depthwise separable convolution ($\mathcal{C} = 1$).

Lighter U-net @$\mathcal{C}$: Both "Conv set 1" and "Conv set 2" correspond to a group convolution with group size $\mathcal{C}$, where $\mathcal{C} \in \{1, 2, 4, 8, 16, 32, 64, 128\}$ and $\mathcal{C} = 1$ represents depthwise separable convolutions. For the first layer with 3 channels, we always set $\mathcal{C}$ to 3 (except when $\mathcal{C} = 1$). For other layers, we choose the minimum value of $\mathcal{C}$ and the input channel number as the final group size.

Note that both "Conv set 1" and "Conv set 2" are applied with Batch Normalization and ReLU activation function. The proposed networks keep the simplicity and elegance of U-net architecture. Their capability for ultrasound images segmentation will be illustrated by the experiments in next section.

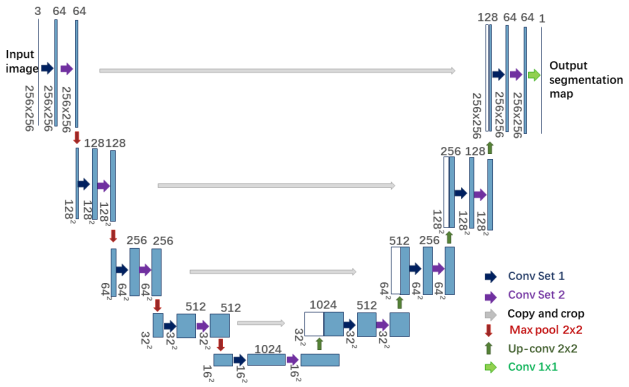

Fig. 1. General architecture of U-net and proposed Light U-net, Lighter U-net @$\mathcal{C}$. The difference between these networks lies in the definition of "Conv set 1" and "Conv set 2".

3.3. Evaluation metrics

We use Dice coefficient (DC) as the main metric and 95% Hausdorff distance (95% HD) as an auxiliary metric to evaluate the segmentation performance of the networks. Note that those are the most common used metrics in medical image segmentation data challenges [5]. All trainings are performed with three different random initializations of model parameters. Besides, data are reshuffled at every epoch and then used for mini-batch training. The mean and standard deviation of our evaluation metrics, computed on test dataset, are reported.

4. EXPERIMENTS

4.1. Comparative analysis

We first summarize the mean and standard deviation of the segmentation performance evaluation metrics of the networks described in Section 3.1, in Table 1. We use the settings fine-tuned on U-net for all the networks (see Section 4.2, for an exhaustive presentation). The only specificity is that the weight decay is tuned separately for each network within the set $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 0\}$. The results show that U-net outperforms all the other architectures, reaching a mean Dice coefficient of 0.929. Note that, although DeepLab v3+ (ResNet-101) and FCN-8s can achieve relatively good results, they all have bigger model sizes than U-net, namely, 54.7 millions parameters for DeepLab v3+ (ResNet-101) and 18.6 millions for FCN-8s, compared to 17.3 millions for U-net.

Method	Backbone	DC	95% HD
DeepLab v3+	MobileNet	0.904±0.003	16.88±0.47
	ResNet-101	0.922±0.003	14.32±0.63
FCN	FCN-32s	0.909±0.005	15.89±0.81
	FCN-16s	0.920±0.003	14.88±0.42
	FCN-8s	0.924±0.001	14.37±0.16
U-net	/	**0.929±0.002**	**13.92±0.34**

Table 1. Segmentation evaluation metrics (mean ± std) of test dataset by different network architectures.

4.2. Hyperparameter tuning strategy for U-net

Let us disclose in detail our hyperparameter tuning process. We only show the mean Dice coefficient of U-net for simplicity of presentation. We first apply the following default setting: No data augmentation, binary cross entropy (BCE) loss, stochastic gradient descent (SGD) optimizer, fixed learning rate $= 10^{-3}$, batch size $= 2$ and weight decay $= 0$. With such setting and training strategy, U-net achieved a mean Dice coefficient of 0.855. Let us now explain how we tuned the hyperparameters step by step to improve the segmentation performance.

Data augmentation: Random flipping, rotating, scaling, translating and shearing are applied to each image as data augmentation skills to restrain overfitting and improve the generation capability. This leads to a mean Dice coefficient improvement from 0.855 to 0.877.

Loss function: Compared to BCE loss, Dice loss (DL) [13] improves the Dice coefficient from 0.877 to 0.891.

Optimizer and learning rate: Adam optimizer [23] with a learning rate of 10^{-4} is used to replace SGD optimizer, yielding Dice coefficient improvement from 0.891 to 0.911. Moreover, adopting the "reduce learning rate on plateau" strategy, i.e. reduce by half the learning rate if the validation loss does not decrease for 5 consecutive epochs, leads to a Dice coefficient improvement from 0.911 to 0.917.

Batch size: The Dice coefficient again improves from 0.917 to 0.928 by adjusting the batch size from 2 to 8.

Weight decay: Weight decay [24, 25] helps to restrain overfitting and improves the generalization capability. In our experiment, the Dice coefficient improves from 0.928 to 0.929 after tuning weight decay from 0 to 10^{-4}.

To conclude, we have achieved an improvement of the mean Dice coefficient of U-net, from 0.855 to 0.929, by tuning the hyperparameters of the network.

[5] See `https://grand-challenge.org/`

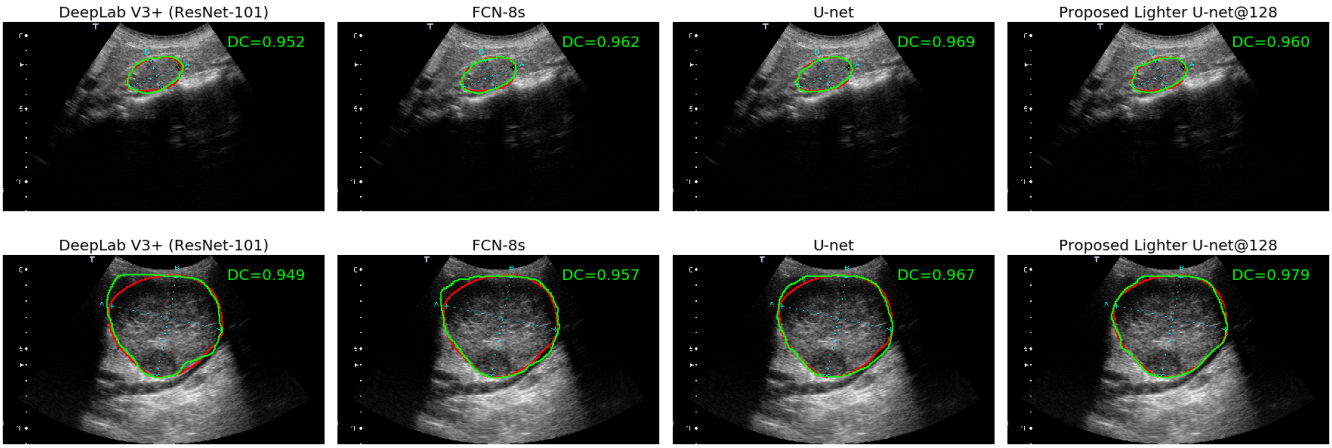

Fig. 2. Visualization of the segmentation results by different networks. The calipers are in light blue color. The contour in red and in green correspond to the ground truth and the predicted lesion segmentation, respectively. First row: lymphadenopathy. Second row: hepatic tumor.

4.3. Experiments of proposed lightweight U-net versions

We then test the proposed Light U-net and Lighter U-net @$\mathcal{C}$ architectures with $\mathcal{C} \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, and the same fine-tuned hyperparameters stated above. Again, the only change lies in the weight decay, tuned separately for each network. The results are reported in Table 2. We also specify the number of parameters to store, for each trained network.

	Parameters	**DC**	**95% HD**
U-net	17,267,393	**0.929±0.002**	**13.92±0.34**
Light U-net	11,000,257	0.927±0.001	14.16±0.47
Lighter U-net @128	7,944,661	**0.928±0.001**	**13.93±0.50**
Lighter U-net @64	4,995,541	0.926±0.001	14.21±0.24
Lighter U-net @32	3,465,685	0.926±0.001	14.51±0.09
Lighter U-net @16	2,700,757	0.922±0.003	15.18±0.43
Lighter U-net @8	2,318,293	0.921±0.002	15.14±0.43
Lighter U-net @4	2,127,061	0.920±0.003	15.22±0.58
Lighter U-net @2	2,031,445	0.920±0.001	15.52±0.19
Lighter U-net @1	1,983,583	0.917±0.001	15.98±0.23

Table 2. Comparison between U-net and its lighweight variants.

First, one can notice that the segmentation performance using Light U-net is slightly damaged, despite of an obvious reduced model size. Regarding Lighter U-net, introducing depthwise separable convolution ($\mathcal{C} = 1$) to U-net can reduce the network size significantly, but it comes along with an obvious decrease in the network segmentation performance. As $\mathcal{C}$ increases, the performance as well as the model size increases gradually. When $\mathcal{C} = 128$, the Lighter U-net @128 network can achieve a mean Dice coefficient 0.928, which is comparable to U-net (mean Dice coefficient of 0.929), but with much smaller model size. This provides us a conjecture that for layers with too many input channels, only parts of the channels are related and should be implemented into simultaneous spatial convolution. In our experiment, using 128 channels for performing the spatial convolutions at a time appears enough for keeping the network capability, along with an obvious reduction in model size. To make the best compromise between model size and segmentation performance, Lighter U-net @128 is chosen as our final lightweight segmentation network.

Figure 2 displays some examples of the lesion segmentation results extracted from the test dataset with different neural networks. We choose here two very different kinds of tumors, namely lymphadenopathy (first row) and hepatic tumor (second row). We can see that for both cases, the networks can provide a quite accurate tumor location. But we must note that the electronic calipers imposed by the radiologists might have influenced the results as they introduce additional information from experts to the data. On the other hand, performance for clearly delineating the tumor boundaries can vary between the networks. In those two examples, one can see that the proposed Lighter U-net @128 has comparable performance with U-net and better performance than FCN-8s and DeepLab v3+ (ResNet-101) to delineate the tumor boundaries.

5. CONCLUSION

In this paper, we perform an extensive comparative study for the segmentation of lesions on a recent multicentric and multipathology ultrasound image dataset. Classic networks such as FCN, DeepLab v3+ and U-net are first compared to fulfill the automatic lesion segmentation. The best performance are reached by U-net with a Dice coefficient of 0.929. To help the reproducibility of the results, we disclose in detail the process of hyperparameter tuning. Furthermore, we introduce group convolution, including depthwise separable convolution as a particular case, to yield lightweight versions of U-net. In particular, the proposed Lighter U-net @128 architecture is shown to achieve comparably good segmentation performance while requiring much less model parameters.

6. COMPLIANCE WITH ETHICAL STANDARDS

The approval of the study was granted by the ethics committee of each involved institution and was declared to the French Commission Nationale Informatique et Liberté (CNIL declaration No. 912346).

7. ACKNOWLEDGMENTS

The PhD thesis of Yingping Li is supported by the China Scholarship Council (CSC).

8. REFERENCES

[1] R. Forghani, P. Savadjiev, A. Chatterjee, and al., "Radiomics and artificial intelligence for biomarker and prediction model development in oncology," *Computational and Structural Biotechnology Journal*, vol. 17, pp. 995, 2019.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 2015, pp. 3431–3440.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[4] L. C. Chen, Y. Zhu, G. Papandreou, and al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV 2018)*, 2018, pp. 801–818.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, 2017, pp. 2961–2969.

[6] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[7] Z. Wang, Z. Zhang, J. Zheng, and al., "Deep learning in medical ultrasound image segmentation: a review," *arXiv preprint arXiv:2002.07703*, 2020.

[8] Y. Zhang, M. T. C. Ying, L. Yang, and al., "Coarse-to-fine stacked fully convolutional nets for lymph node segmentation in ultrasound images," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2016)*. IEEE, 2016, pp. 443–448.

[9] H. Chen, Q. Dou, D. Ni, and al., "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 507–514.

[10] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *arXiv preprint arXiv:1611.08408*, 2016.

[11] J. Lee, E. Kim, S. Lee, and al., "Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, 2019, pp. 5267–5276.

[12] L. Wang, K. Lekadir, S. L. Lee, and al., "A general framework for context-specific image segmentation using reinforcement learning," *IEEE transactions on medical imaging*, vol. 32, no. 5, pp. 943–956, 2013.

[13] S. A. Taghanaki, K. Abhishek, J. P. Cohen, and al., "Deep semantic segmentation of natural and medical images: A review," *Artificial Intelligence Review*, pp. 1–42, 2020.

[14] N. Lassau, L. Chapotot, B. Benatsou, and al., "Standardization of dynamic contrast-enhanced ultrasound for the evaluation of antiangiogenic therapies: the french multicenter support for innovative and expensive techniques study," *Investigative Radiology*, vol. 47, no. 12, pp. 711–716, 2012.

[15] N. Lassau, J. Bonastre, M. Kind, and al., "Validation of dynamic contrast-enhanced ultrasound in predicting outcomes of antiangiogenic therapy for solid tumors: the french multicenter support for innovative and expensive techniques study," *Investigative Radiology*, vol. 49, no. 12, pp. 794, 2014.

[16] K. He, X.u Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] A. G. Howard, M. Zhu, B. Chen, and al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017, pp. 1251–1258.

[21] N. K. Jha, R. Saini, S. Nag, and S. Mittal, "E2gc: Energy-efficient group convolution in deep neural networks," in *Proceedings of the 33rd International Conference on VLSI Design and Embedded Systems (VLSID 2020)*. IEEE, 2020, pp. 155–160.

[22] L. Sifre, *Rigid-motion scattering for image classification*, Ph.D. thesis, Ecole Polytechnique, CMAP, 6 Oct. 2014.

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

[24] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proceedings of Advances in Neural Information Processing Systems (NIPS 1992)*, 1992, pp. 950–957.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.