

Performance evaluation of lightweight convolutional neural networks on retinal lesion segmentation

M. Siebert^a and P. Rostalski^{a,b}

^aInstitute for Electrical Engineering in Medicine, University of Lübeck, Lübeck, Germany

^bFraunhofer IMTE, Lübeck, Germany

ABSTRACT

In addition to the recent development of deep learning-based, automatic detection systems for diabetic retinopathy (DR), efforts are being made to integrate those systems into mobile detection devices running on the edge requiring lightweight algorithms. Moreover, to enable clinical deployment it is important to enhance the transparency of the deep learning systems usually being black-box models and hence giving no insights into its reasoning. By providing precise segmentation masks for lesions being related to the severity of DR, a good intuition about the decision making of the diagnosing system can be given. Hence, to enable transparent mobile DR detection devices simultaneously segmenting disease-related lesions and running on the edge, lightweight models capable to produce fine-grained segmentation masks are required contradicting the generally high complexity of fully convolutional architectures used for image segmentation. In this paper, we evaluate both the runtime and segmentation performance of several lightweight fully convolutional networks for DR related lesion segmentation and assess its potential to extend mobile DR-grading systems for improved transparency. To this end, the U²-Net is downscaled to reduce the computational load by reducing feature size and applying depthwise separable convolutions and evaluated using deep model ensembling as well as single- and multi-task inference to improve performance and further reduce memory cost. Experimental results using the U²-Net-S[†] ensemble show good segmentation performance while maintaining a small memory footprint as well as reasonable inference speed and thus indicate a promising first step towards a holistic mobile diagnostic system providing both precise lesion segmentation and DR-grading.

Keywords: mobile segmentation, diabetic retinopathy, deep learning, multi-lesion segmentation, U-Net, fundus image

1. INTRODUCTION

Diabetic retinopathy (DR) is a pathologic condition related to diabetic changes of the vascular tissue causing retinal lesions such as microaneurysms, haemorrhages, hard- and soft exudates,¹ as shown in fig. 1. Later on vascular closure, hypoxia, vascular proliferation and retinal detachment can occur finally resulting in heavy impairment or total loss of vision when being untreated.¹ These changes become visible in the eye's fundus which enables trained specialists, i.e. ophthalmologists, to screen diabetic patients for early signs of DR and, when present, prescribe therapeutic measures to prevent the progression of DR to vision-threatening stages.¹

Leveraging automatic diagnostic algorithms to detect the presence and severity of DR, could help general practitioners in screening diabetic patients in rural areas where access to specialized medical examination is limited.² To solve this task, deep learning assisted DR detection systems are of high research interest due to the excellent diagnostic performance that can be achieved using neural networks. To additionally evade the high cost of bulky specialized hardware that is able to run state-of-the-art but large deep learning models, developing lightweight, mobile, edge device applicable systems with good performance gains increasing attention as deployment of those is less expensive and can improve usability and availability in areas with limited access to medical care.

Nevertheless, deep learning systems are usually black-box models that lack transparency limiting the trust in DL algorithms and the expressive power of predicted diagnoses. Thus, providing insights into the decision making and visualizing intermediate results can be helpful for clinicians to verify the model's prediction. In

Send correspondence to M. Siebert: m.siebert@uni-luebeck.de

M. Siebert and P. Rostalski, "Performance evaluation of lightweight convolutional neural networks on retinal lesion segmentation" in *Medical Imaging 2022: Computer-Aided Diagnosis*, K. Drukker and K. M. Iftikharuddin and H. Lu and M. A. Mazurowski and C. Muramatsu and R. K. Samala, no. 12033, pp. 806 - 817, SPIE, 2022; <https://doi.org/10.1117/12.2611796>.

Copyright 2022 Society of PhotoOptical Instrumentation Engineers. One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

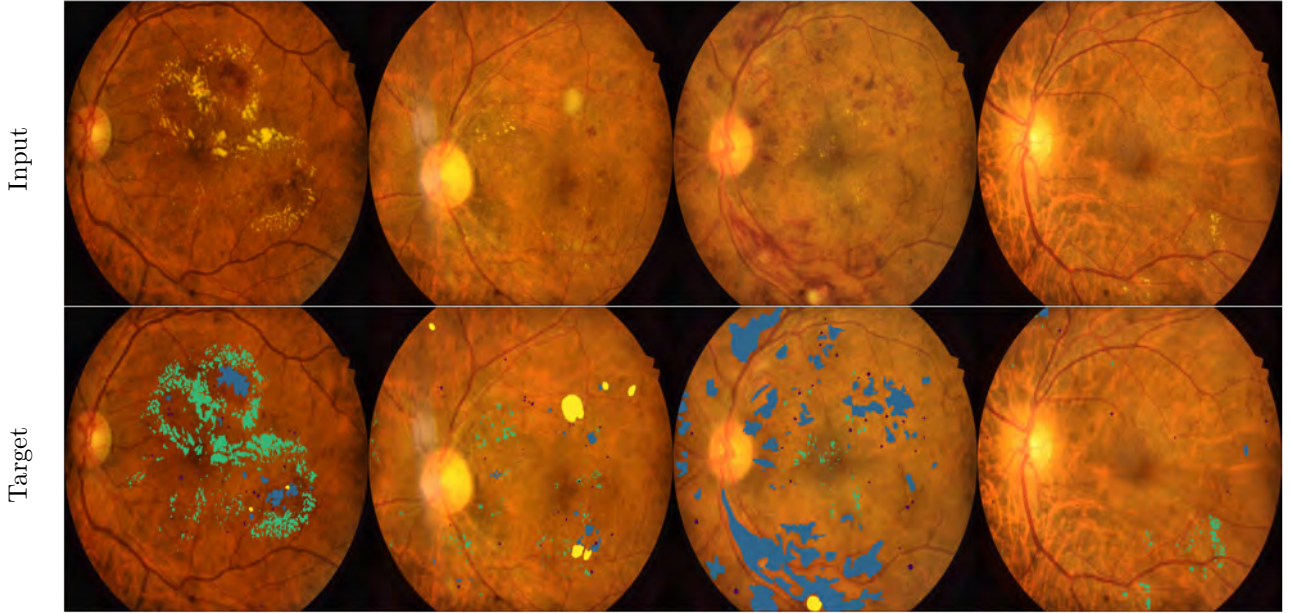


Figure 1. Sample images of the IDRiD dataset in the first row with corresponding segmented lesions in the second row (microaneurysms ●, haemorrhages ●, hard exudates ●, soft exudates ●).

particular, highlighting retinal lesions associated with the vision-threatening disease at hand and present at the patient’s retina can improve the physician’s understanding of the deep learning system’s prediction of whether the disease is present and how reliable this prediction is. To this end, the DL grading system can be accompanied by or even build based on a segmentation model predicting fine-grained masks of DR related lesions. However, these models are typically more computationally demanding compared to regular classification models due to the large intermediate activation maps computed e.g. in the decoder part of the U-Net.

Therefore, we apply various complex instances of the U²-Net,³ a recent extension to the famous U-Net,⁴ for the specific task of Diabetic Retinopathy (DR) related lesion segmentation and analyse its suitability for edge-device implementation in this work. According to the authors, the U²-Net shows state-of-the-art performance for salient object detection when trained from scratch, while being designed to maintain low memory cost and keep input image resolution high. Hence, the U²-Net promises to be a good baseline for building a lightweight segmentation model being applicable to edge devices and generalizing well on medical image data that, in particular for the task of lesion segmentation, is commonly rare. By applying mobile application-suitable convolutions, i.e. depthwise separable convolutions,⁵ varying the network capacity as well as evaluating the impact of multi-task and ensemble learning, we aim to evaluate (a) the suitability of the U²-Net to the task of biomedical image segmentation, (b) the segmentation performance when training the U²-Net on small data from scratch and (c) the segmentation accuracy and computational load of the differently complex instances of the model and hence the suitability of the architecture for edge device implementation. To this end, we compare the performance of the basic U-Net to the U²-Net and state-of-the-art results in the literature on the one hand. On the other hand, we benchmark the different model instances to find a good trade-off between computational cost and segmentation performance of the retinal lesions.

2. RELATED WORK

With the rising performance of deep learning algorithms over the last two decades, developing artificial intelligence-based automatic medical diagnostic algorithms gained tremendous research interest. This applies to the task of DR detection⁶⁻¹¹ as well, leading to the first algorithms being approved for use in clinical care.^{12,13} In addition to that, also mobile DR detection systems¹⁴ using lightweight neural networks¹⁵ are brought into research focus.

Despite their excellent performance, deep learning algorithms are not guaranteed to generalize to real-world application data and are black-boxes that in general do not allow to understand the model’s reasoning. This however is inconvenient for highly safety-critical applications such as medical diagnosis where it is important to understand why a prediction was made, to verify the decision and to prevent severe misdiagnoses. Therefore, recent methods are built to highlight areas in the retinal fundus that contributed to the model’s prediction the most^{7,8,14,16} or to grade the disease based on detected DR related lesions.^{9,10,12}

Additionally, computing precise segmentation masks of lesion-specific pathologic areas in the retina could give further insights and hence improve classification. This, however, is a highly demanding task due to the lack of a sufficient amount of annotated data as well as the lesions being commonly very small and also sometimes poorly distinguishable from noise and other retinal structures and lesions. To circumvent the problem of data availability for DR lesion segmentation tasks, semi- or weakly-supervised as well as adversarial training strategies can be used to leverage information of image-level graded datasets,^{17–19} such as Kaggle DR²⁰ or Messidor.²¹ Accounting for the small lesion size of e.g. microaneurysm and hard exudate, high-resolution input images and multi-scale approaches can boost segmentation performance fusing important information from local and global image features into the model’s prediction.^{22–24} To additionally account for edge device suitable segmentation, Guo et al.²⁵ propose the LWENet architecture with about 1.9M parameters for lightweight hard exudate segmentation using only a single pooling operation, running at approximately 11.1 fps for an input image size of (1440 × 960) pixels.

3. METHODS

To combine both, high segmentation performance and low computational cost while training on a small medical image data set from scratch, this work implements the U²-Net with different model capacities and analyses its suitability for DR lesion segmentation with respect to edge device implementation.

As a reference model, the U-Net is used as baseline architecture being an encoder-decoder shaped fully convolutional network explicitly designed for biomedical image segmentation on a small amount of data.⁴ The encoder is trained to extract meaningful features using sequential convolution, batch-norm, ReLU operations while scaling down the input image resolution using max-pooling operations and hence increasing the receptive field of the model, similar to most classification architectures. However, mirroring the encoder structure, the decoder reconstructs fine-grained segmentation masks from the learned latent encoding by using skip connections to exploit high-resolution features from the intermediate layers in the encoder. This enables the U-Net to use both global and local features to produce high resolution and precise segmentation masks in a single forward pass of the network requiring only few images, which is beneficial for the task of biomedical image segmentation where data is usually scarce.⁴ For this work, we use the original U-Net architecture comprising five depth layers but with additional batch-norm layers and transposed convolutions in the decoder.

Adopting the encoder-decoder structure with additional skip connections, Qin et al.³ propose the U²-Net that makes use of a nested U-Net structure using so-called, residual U-blocks (RSU) which exploit residual connections as well as pooling operations and dilated convolutions to increase the model depth and enhance the receptive field of each depth-level in the wrapping U-Net. With this, the RSU-block extracts rich multi-scale features already at early layers on both local and global level while retaining high resolution of the input activation maps leading to an increased model capacity and performance of the U²-Net compared to the plain U-Net.³ Despite increasing the number of model parameters, Qin et al. show that using the RSU-block introduces only a small increase in computational cost compared to the plain U-Net and has significantly lower cost in contrast to other frequently used U-Net extensions, e.g. residual or dense blocks while performing similar or better to other state-of-the-art salient object detection models. To further improve model performance, Qin et al. additionally apply deep supervision to side outputs of each depth-level within the decoder. As a result, the authors claim a good inference performance with using the U²-Net preserving fine-grained structures, enabling training on small data from scratch and alleviating the design of lightweight, mobile architectures due to being independent of large pretrained backbones and comparably small computational load.

In this work, we first implement the two models introduced by Qin et al. without any changes, here referred to as U²-Net-O and U²-Net-M, and analyse their performance on biomedical image segmentation data in comparison

Table 1. Setup of differently scaled U²-Nets. The first and second column displays the number of input and middle channels of each RSU-block ($e_i, i = 1, \dots, 6$) in the encoder, respectively. The decoder is adopted such to the number of features in the encoder is mirrored. The third row displays the number of features of the final convolutional layer computing the high-resolution output segmentation mask.

Model	Input						Middle						Output
	e_1	e_2	e_3	e_4	e_5	e_6	e_1	e_2	e_3	e_4	e_5	e_6	
U ² -Net-O	3	64	128	256	512	512	16	32	64	128	256	256	64
U ² -Net-M	3	64	64	64	64	64	16	16	16	16	16	16	64
U ² -Net-S	3	32	32	32	32	32	8	8	8	8	8	8	32
U ² -Net-XS	3	16	16	16	16	16	4	4	4	4	4	4	16

to the baseline U-Net. Both networks apply five consecutive RSU-blocks with subsequent max-pooling in the encoder as well as an additional RSU-block at the model’s bottleneck and only differ in the number of features and hence model size and capacity. We then additionally scale down the U²-Net by reducing the convolution layer’s feature size of the U²-Net-M by 50 % and 75 % resulting in two smaller model instances, i.e. U²-Net-S and U²-Net-XS. A detailed description of the model parameterization is given in table 1.

Moreover, we deploy four additional models using depth-wise separable convolutions. This specific convolutional layer factorizes a regular convolution into two separate operations, i.e. a depth- and a point-wise convolution. While the former processes the input spatially, i.e. applies a single filter for each or a group of channels, the latter uses a 1×1 convolution to expand or reduce the channel depth of the depth-wise convolution. This lowers the number of parameters, memory footprint and required Multiply-Accumulation (MAC) operations significantly while retaining a large quantity of the convolutional capacity compared to regular convolutional layers. Hence, depth-wise separable convolutions are frequently used, e.g. in the MobileNet⁵ and its subsequent versions, to reduce the computational cost of a model architecture to enable edge device application while keeping high performance. To evaluate if this sparse convolution can be beneficial for further reduction of the U²-Net’s computational load while maintaining sufficient performance, we interchange every but each RSU-blocks first and the final U²-Nets output convolution within the differently sized models with depthwise separable convolutions. We refer to these models as U²-Net-O*, U²-Net-M*, U²-Net-S* and U²-Net-XS*.

In addition, we evaluate the effect of model ensembling for every but the largest U²-Net-O, denoted as U²-Net-M[†], U²-Net-S[†], and U²-Net-XS[†], as deep ensembling is known to significantly improve model performance. Moreover, performing dual- (DT) and multi-task (MT) lesion segmentation can help to further decrease memory and computational cost. Therefore, we accordingly set up multi-lesion segmentation models in addition to the single-task (ST) instances. For dual-task training, microaneurysms and haemorrhages, as well as hard- and soft exudates are segmented simultaneously, while during multi-task training a single model learns to segment all four lesions at once. With this, inference time and the overall number of model parameters as well as the computational cost for model training can be significantly reduced by half and a quarter, respectively. However, hard parameter sharing between competing tasks is more difficult to train and can impair model performance.

In table 2, the different single-task model architectures are compared to each other with respect to their computational demands at inference time. To assess the suitability of the networks for mobile implementation, we report the frames per second (fps) the model takes during inference, as well as the computational complexity given as multiply-accumulates (MACs), the number of parameters, the network storage size for a single-precision model and the RAM allocated on the GPU by PyTorch during inference. Please note that dual- and multi-task models are omitted as they do only differ slightly from the single-task model computational cost. However, when using single-task models to segment all four lesions, the reported computational cost is quadrupling and doubling for dual-task models.

4. IMAGE DATA AND PREPROCESSING

For training, the publicly available Indian Diabetic Retinopathy image Dataset²⁶ (IDRiD) is used. It provides 81 retinal fundus images with fine-grained segmentation masks of microaneurysms (MA), haemorrhages (HE)

Table 2. Runtime performance of the implemented single-task networks at inference for a single lesion running on a GeForce RTX 2080Ti (11 GB) with an Intel Xenon Gold 6212U (192 GB) using PyTorch (v1.8.0) and a randomly sampled input image of size $(1 \times 3 \times 512 \times 512)$ for 500 iterations. The MACs of the models are approximated using the toolbox *thop*[†].

Model	Frame rate [fps]	MACs [G]	Parameters [M]	Model size (fp32) [MB]	GPU-RAM [MB]
U-Net	30.46	218.85	31.04	124.17	586.28
U ² -Net-O	27.71	141.98	43.87	175.48	798.34
U ² -Net-O*	31.90	81.20	17.66	70.65	692.76
U ² -Net-M	43.75	51.32	1.13	4.52	579.36
U ² -Net-M*	51.22	33.48	0.66	2.63	576.50
U ² -Net-S	64.85	13.19	0.29	1.14	289.80
U ² -Net-S*	54.24	8.93	0.18	0.70	290.48
U ² -Net-XS	69.26	3.48	0.07	0.29	147.45
U ² -Net-XS*	55.10	2.52	0.05	0.20	147.43
U ² -Net-M [†]	14.40	152.70	3.39	13.57	610.15
U ² -Net-S [†]	21.55	38.91	0.86	3.43	314.86
U ² -Net-XS [†]	22.94	10.10	0.22	0.88	170.84

and hard- (HX) as well as soft exudates (SX) that are important biomarkers for detection of DR in particular in early stages, as displayed in fig. 1. We adhere to the predefined train-test split of the dataset while 20% of the training data are randomly selected at the start of each training and used for performance validation.

The image preprocessing applied in this work comprises cropping the images to the visible retinal disc with a semi-automatic algorithm using binary thresholding, resizing the images to (512×512) pixels and applying contrast limited histogram equalization (CLAHE). The CLAHE operation is applied to the value channel of the input image after converting it to the HSV colour space to equally enhance the contrast of all image colour channels. After converting the images back into RGB colour space, the images are normalized with the mean and standard deviation of the test set.

At training time, data augmentation, i.e. random affine transformations as well as horizontal and vertical flipping, is applied to the input images to artificially scale up the training data set and reduce model overfitting.

5. EXPERIMENTAL SETUP

Each of the above-described models is trained ten times on the IDRID dataset without pretraining. Training was conducted on an Nvidia Tesla-V100 (32 GB) along with an Intel(R) Xeon(R) Platinum 8168 CPU over 400 epochs that takes about 1.5 h per model in single-task mode for the large U²-Net-O and about 1 h for the small U²-Net-XS. We use Adam as optimizer with a learning rate of $5e - 4$, L2-norm weight decay set to $1e - 4$ and a plateau scheduler with a factor of 0.75 and patience of 15 epochs. As loss, a weighted fusion of Dice-loss (\mathcal{L}_D) and focal binary cross-entropy²⁷ ($\mathcal{L}_{\text{FBCE}}$) is applied to the model according to

$$\mathcal{L}(y, \hat{y}) = \frac{1}{C} \sum_k^C \left(\alpha \cdot \mathcal{L}_{\text{FBCE}}(y_k, \hat{y}_k, \gamma) + (1 - \alpha) \cdot \mathcal{L}_D(y_k, \hat{y}_k) \right) \cdot w_k$$

with w_k being a weighting factor only active during dual- and multi-task training boosting the learning of under-represented classes with low volume in the current batch. Accordingly, w_k is set to zero in case a class is absent throughout the batch to prevent the model from degrading to constantly predict empty masks. This mainly affects soft exudate segmentation, as this lesion is only present in about half the images of the IDRID data set. We chose $\alpha = 0.75$ prioritizing the focal binary cross-entropy to smooth the Dice-loss but still benefit from the

[†]<https://github.com/Lyken17/pytorch-OpCounter> (last accessed: 02/16/2022)

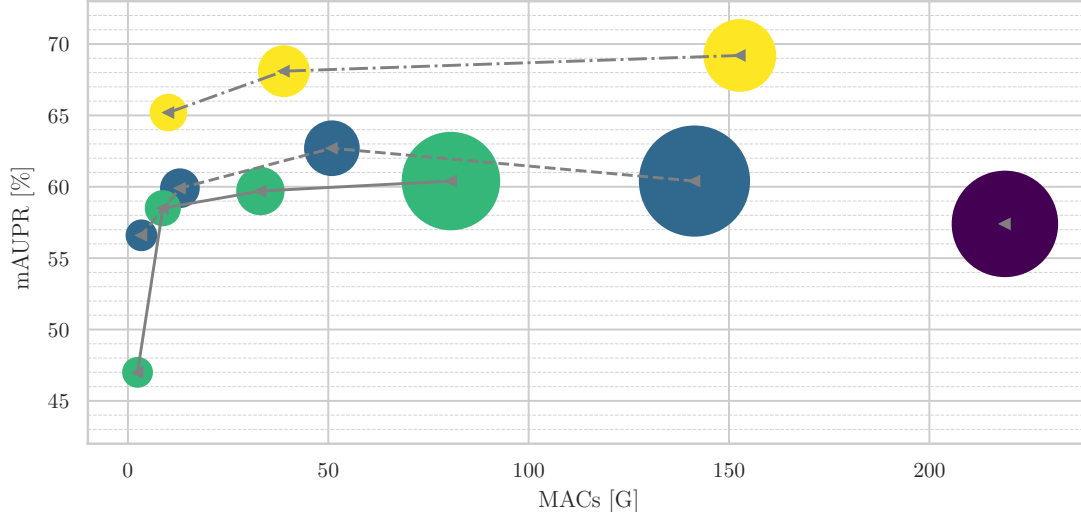


Figure 2. Comparison of the U-Net (purple: ●), the differently scaled U²-Nets (blue: ●), their counterparts with depthwise separable convolutions (green: ●) and the ensembles U²-Net-XS[†], U²-Net-S[†] and U²-Net-M[†] (left to right, yellow: ●) with respect to the mean AUPR (mAUAPR) across all lesions and the computational cost in MACs in single-task mode. Circle size corresponds to the model’s number of parameters.

letter’s insensitivity to the class imbalance. Furthermore, we set $\gamma = 1.0$ for the focal binary cross-entropy forcing the model to focus on difficult pixels rather than optimizing easy decisions. Moreover, deep supervision at every depth layer of the network with equally weighted errors of all side outputs of the U²-Net is used to optimize the model following the implementation of Qin et al.³ To this end, the binary target masks are downscaled to the output resolution of each depth layer before evaluating the loss. In addition to the class-dependent weighting for dual- and multi-task training, the number of epochs and the patience of the scheduler are increased to 600 and 800 as well as 30 and 50, respectively, to ensure model convergence.

At test time, we measure the performance of the predicted segmentation masks in terms of Dice-score (DS), Area under the Precision-Recall Curve (AUPR) and Hausdorff-distance of the predicted segmentation to the target mask. The threshold for computing binary segmentation masks and with this both the Dice-score and the Hausdorff-distance is selected from the Precision-Recall curve to maximize the harmonic mean of both precision and recall and hence the dice-score for each class individually. We include measuring the Hausdorff-distance, which is computed using the Toolbox provided by DeepMind[‡], as the Dice-Score is very sensitive to small changes when dealing with only small lesions which is the case for microaneurysms and hard exudates. We set the surface distance to the maximum value corresponding to the image size when no lesion is present but the network’s prediction is not empty and vice versa, which is only affecting soft exudate segmentation as the other lesions are present in every image of the test set.

6. RESULTS AND DISCUSSION

The results of this work are presented in fig. 2, showing the mean AUPR score with respect to the tested model’s computational complexity as well as the number of parameters represented by their corresponding circle size. Furthermore, in table 3 the lesion-specific segmentation AUPR score and a comparison to other state-of-the-art models are presented. More detailed results of all ten individual training runs are provided in form of boxplots in figs. 3 to 5 showing the lesion-specific AUPR, Dice-score and Hausdorff-distance, for both the regular and depth-wise separable convolutional U²-Net models.

Having significantly more parameters compared to the U-Net and an additional depth layer, it is visible from fig. 2 that the U²-Net-O expectedly outperforms the U-Net on the lesion segmentation task by 5.2% with respect

[‡]<https://github.com/deepmind/surface-distance> (last accessed: 01/23/2022)

Table 3. The results on the IDRiD test data set for the different single-task models (if not otherwise stated) are given as mean area under Precision-Recall-curve (AUPR) across the individual training runs per lesion excluding outliers with the standard deviation enclosed in brackets and the overall mean AUPR (mAUPR) across all four lesions.

Model	MA	HE	HX	SX	mAUPR
U-Net	0.355 (0.010)	0.555 (0.040)	0.747 (0.052)	0.639 (0.088)	0.574
U ² -Net-O	0.309 (0.111)	0.605 (0.036)	0.766 (0.063)	0.737 (0.030)	0.604
U ² -Net-O*	0.343 (0.012)	0.637 (0.020)	0.731 (0.062)	0.705 (0.022)	0.604
U ² -Net-M	0.345 (0.030)	0.648 (0.015)	0.790 (0.048)	0.724 (0.045)	0.627
U ² -Net-M*	0.318 (0.085)	0.622 (0.034)	0.730 (0.057)	0.716 (0.040)	0.597
U ² -Net-S	0.345 (0.023)	0.606 (0.070)	0.760 (0.057)	0.686 (0.057)	0.599
U ² -Net-S*	0.331 (0.030)	0.575 (0.037)	0.775 (0.064)	0.658 (0.049)	0.585
U ² -Net-XS	0.318 (0.036)	0.558 (0.043)	0.774 (0.057)	0.615 (0.070)	0.566
U ² -Net-XS*	0.209 (0.102)	0.418 (0.160)	0.701 (0.079)	0.552 (0.119)	0.470
U ² -Net-S (DT)	0.370 (0.017)	0.632 (0.016)	0.753 (0.044)	0.697 (0.038)	0.613
U ² -Net-S (MT)	0.349 (0.018)	0.510 (0.034)	0.723 (0.061)	0.707 (0.026)	0.572
U ² -Net-M [†]	0.407	0.694	0.862	0.805	0.692
U ² -Net-S [†]	0.402	0.678	0.850	0.795	0.681
U ² -Net-XS [†]	0.375	0.639	0.852	0.743	0.652
Zhou et al. ¹⁸	0.496	0.694	0.887	0.741	0.704
Yan et al. ²²	0.525	0.703	0.889	0.679	0.699
Guo et al. ²⁴	0.463	0.637	0.795	0.711	0.652

to the mAUPR score. Thereby, the most improvement is visible for haemorrhage and soft exudate segmentation, while hard exudate segmentation is only slightly better and microaneurysm segmentation even performs worse according to table 2. However, the U²-Net-O saves about 35.4% of MACs and hence only decreases the frame rate slightly by about 9% from 30.46 fps to 27.71 fps showing the high computational efficiency of the U²-Net’s architectural design. Unexpectedly, we observe the U²-Net-M to outperform the U²-Net-O except for soft exudate segmentation despite the reduced model size and capacity. From fig. 3 it is visible that this mainly originates from overfitting which is occurring for the haemorrhage segmentation using the largest U²-Net. However, further downscaling the model, i.e. using U²-Net-S and U²-Net-XS, decreases the measured performance as expected. With this, the mean AUPR of the U²-Net-XS is on average on par with the U-Net, however reducing model size from 31.04 M to 0.07 M parameters and more than doubling the frame rate.

Additionally, the hard exudate segmentation AUPR- and Dice-score show an overall high variance across the results. This in contrast is not visible for the measured Hausdorff-distance, presented in fig. 5, showing increasingly good performance for hard exudate segmentation with growing U²-Net model capacity, underlining the high sensitivity of the AUPR and Dice-score to small pixel level variations. This, however, do not affect the lesion detection performance on instance level measured through Hausdorff-distance, indicating that measuring segmentation performance solely through pixel-level metrics may not be optimal. Hence, looking at the Hausdorff-distance for haemorrhage segmentation the decrease in performance for the larger U²-Nets, visible in the AUPR score, is reduced, suggesting the overfitting to occur on pixel level. Although, computing Hausdorff-distance for soft exudates is also not ideal due to the heavy penalty when no lesion is present and only a single pixel is segmented by the model which leads to an overall high and noisy Hausdorff-distance.

Moreover, with using depthwise separable convolutions model performance expectedly is at most on par or decreases slightly compared to the neural networks with regular convolutions due to the reduced model capacity. While the decrease of the computational cost has a beneficial impact for the original sized U²-Net-O as presented in table 2, the effect significantly diminishes for the smaller models. Exemplary, using depthwise separable convolutions only marginally decreases the model size as well as required MAC-operations and even increases inference frame rate comparing the U²-Net-XS to U²-Net-XS*. From this, we reason that in the low size regime of the tested architectures the reduced computational load using depthwise separable convolutions

does not compensate for the loss of performance. Nevertheless, this may be attributed to some extent to the fact that the usage of depthwise separable convolutions is not fully optimized using PyTorch²⁸ and thus also could be different for more optimized toolboxes.

Overall, we reason that the U²-Net-S provides the best trade-off between performance and computational efficiency when using individual single-task model instances. However, an ensemble of the three top-performing U²-Net-XS models outperforms the former with respect to GMACs (−22.1%), model size (−24.1%) and mAUPR (+8.9%) significantly. Equally, by using the ensembles U²-Net-S[†] and U²-Net-M[†] we observe a further increase in model performance by +9.6% and +14.6% compared to the U²-Net-M and U²-Net-O, respectively, while having less or similarly MAC-operations and parameters. Although providing the overall best performance, all the ensemble models have very high inference runtime due to being serially executed in our implementation. This, however, could be eliminated by parallelizing the model inference by exploiting grouped convolutions²⁹ if possible which on the downside would increase memory usage significantly. Referring to fig. 2, we propose to use an ensemble of either the U²-Net-XS or U²-Net-S that provide the best trade-off between model performance and computational and hardware requirements.

Furthermore comparing dual- and multi-task training, the former yields slightly better performance to the single-task mode, in particular for microaneurysm and haemorrhage segmentation, while the latter visibly suffers from the hard parameter sharing impairing haemorrhage and hard exudate segmentation, as visible from table 3. As a result, using dual-task training promises a reasonable method to reduce the computational load without significant loss of performance and is estimated to further increase the ratio between performance and computational cost being used in an ensemble model.

From the comparison to the literature in table 3 and the official subchallenge-1 leaderboard,³⁰ it is visible that using the U²-Net architecture, in particular with using model ensembling, is achieving state-of-the-art performance for soft exudate segmentation and is on par for hard exudate and haemorrhage segmentation but not reaching the performance for microaneurysm segmentation. The latter is estimated to be caused by the deep network structure preventing good convergence and the comparable low-resolution input image resulting in a loss of essential information as microaneurysm lesions are typically only a few pixels wide. In comparison, the L-Seg model proposed by Guo et al.,²⁴ being based on a pretrained VGG16 backbone with additional side outputs allowing for multi-lesion segmentation, is trained on high-resolution input images with (1440 × 960) pixels and with this outperforms the U²-Net for microaneurysm segmentation. Similarly, Yan et al.²² proposed a Dual-U-Net architecture comprising a fusion of a global U-Net predicting on the complete but downsampled retinal image, a local U-Net operating on high-resolution image patches. In their paper, they show that this Dual-U-Net architecture is boosting segmentation performance of small lesions like microaneurysms and hard exudates. However, their results indicate that larger and more homogenous lesions, i.e. haemorrhages and soft exudates are segmented best using only the global network. Moreover, Zhou et al.¹⁸ deploy model training with (640 × 640) sized fundus images and additionally exploit image-level graded data in a complex adversarial training strategy with weak supervision to improve multi-task DR lesion segmentation performance. In contrast to our findings, they show that using an Xception³¹-based U-Net architecture, which uses depthwise separable convolutions in a more sophisticated manner, could indeed improve the segmentation performance over a plain U-Net implementation using only regular convolutional layers.

However compared to the LWENet,²⁵ the U²-Net-S is more than six times smaller and similarly fast on the same image resolution. Despite the Dice-score of the LWENet being significantly higher as for the U²-Net-S when pretrained on the DDR³² data set (DS=0.782% vs. DS=0.705%), the LWENet performs slightly worse than the U²-Net-S when no pretraining is applied (DS=0.697%). Moreover, using the U²-Net-XS[†] ensemble achieves similar performance (DS=0.779%) compared to the pretrained LWENet without exploiting transfer learning itself and still being about six times smaller aligning the observations of Qin et al.³

7. CONCLUSION

From the presented results we conclude that (a) the U²-Net is an overall suitable architecture for biomedical image segmentation, (b) the model architecture shows good performance even when no pretraining is conducted, that is on par with and in particular for soft exudate segmentation outperforms current state-of-the-art results

using model ensembling (c) the downscaled U²-Net may be a suitable method for edge device implementation as it shows very promising results in fine-grained lesion segmentation while having very few parameters and hence low memory cost as well as a reasonable computational complexity for the task of screening retinal images for lesions.

However, improving overall model performance, in particular for microaneurysm segmentation by using e.g. multi-scale approaches, will be subject to future research. In addition to the already high reduction of model size, further optimization of the computational load to performance trade-off for edge device deployment could be achieved exploiting dual- and multi-task training as well as using more refined mobile building blocks, e.g. Xception-,³¹ Squeeze-and-Excitation-³³ or MBConv-blocks,³⁴ and exploiting model-pruning and distilling. Also, applying more sophisticated training strategies, i.e. adversarial as well as self- or weakly supervised training, potentially can enhance model performance and help implement a transparent, holistic DR detection system comprising both disease grading and lesion segmentation for use on edge devices.

ACKNOWLEDGMENTS

This research is part of the project “Patientennahe Smartphone-basierte Diagnostik” (PASBADIA) kindly supported by the *Joachim Herz Foundation*, Hamburg, Germany.

REFERENCES

- [1] Fong, D. S., Aiello, L., Gardner, T. W., King, G. L., Blankenship, G., Cavallerano, J. D., Ferris, F. L., and Klein, R., “Retinopathy in Diabetes,” (jan 2004).
- [2] Ting, D. S. W., Cheung, G. C. M., and Wong, T. Y., “Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review,” (may 2016).
- [3] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M., “U²-Net: Going Deeper with Nested U-Structure for Salient Object Detection,” *Pattern Recognition* (may 2020).
- [4] Ronneberger, O., Fischer, P., and Brox, T., “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in [*Medical Image Computing and Computer-Assisted Intervention (MICCAI)*], **LNCS 9351**, 234–241, Springer (2015).
- [5] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” (apr 2017).
- [6] Foo, A., Hsu, W., Lee, M. L., Lim, G., and Wong, T. Y., “Multi-task learning for diabetic retinopathy grading and lesion segmentation,” in [*Proceedings of the 32nd Innovative Applications of Artificial Intelligence Conference, IAAI 2020*], **34**, 13267–13272 (apr 2020).
- [7] Gargeya, R. and Leng, T., “Automated Identification of Diabetic Retinopathy Using Deep Learning,” *Ophthalmology* **124**, 962–969 (jul 2017).
- [8] Quellec, G., Charrière, K., Boudi, Y., Cochener, B., and Lamard, M., “Deep image mining for diabetic retinopathy screening,” *Medical Image Analysis* **39**, 178–193 (2017).
- [9] Wang, Z., Yin, Y., Shi, J., Fang, W., Li, H., and Wang, X., “Zoom-in-net: Deep mining lesions for diabetic retinopathy detection,” in [*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*], **10435 LNCS**, 267–275 (2017).
- [10] Yang, Y., Li, T., Li, W., Wu, H., Fan, W., and Zhang, W., “Lesion detection and grading of diabetic retinopathy via two-stages deep convolutional neural networks,” in [*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*], **10435 LNCS**, 533–540 (2017).
- [11] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., and Webster, D. R., “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA* **316**, 2402 (dec 2016).
- [12] Abràmoff, M. D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J. C., and Niemeijer, M., “Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning,” *Investigative Ophthalmology & Visual Science* **57**, 5200 (oct 2016).

- [13] Bhaskaranand, M., Ramachandra, C., Bhat, S., Cuadros, J., Nittala, M. G., Sadda, S. R., and Solanki, K., “The value of automated diabetic retinopathy screening with the EyeArt system: A study of more than 100,000 consecutive encounters from people with diabetes,” *Diabetes Technology and Therapeutics* **21**, 635–643 (nov 2019).
- [14] Natarajan, S., Jain, A., Krishnan, R., Rogy, A., and Sivaprasad, S., “Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening With an Offline Artificial Intelligence System on a Smartphone,” *JAMA Ophthalmology* **137**, 1182 (oct 2019).
- [15] Sheikh, S. and Qidwai, U., “Using MobileNetV2 to Classify the Severity of Diabetic Retinopathy,” *International Journal of Simulation Systems Science & Technology* (mar 2020).
- [16] Gondal, W. M., Kohler, J. M., Grzeszick, R., Fink, G. A., and Hirsch, M., “Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images,” in [*Proceedings - International Conference on Image Processing, ICIP*], **2017-Septe**, 2069–2073 (2018).
- [17] Playout, C., Duval, R., and Cheriet, F., “A Novel Weakly Supervised Multitask Architecture for Retinal Lesions Segmentation on Fundus Images,” *IEEE Transactions on Medical Imaging* **38**, 2434–2444 (oct 2019).
- [18] Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., and Shao, L., “Collaborative learning of semi-supervised segmentation and classification for medical images,” in [*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*], **2019-June**, 2074–2083, IEEE (jun 2019).
- [19] Zhou, Y., Wang, B., Huang, L., Cui, S., and Shao, L., “A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading, and Transferability,” *arXiv*, 1–11 (aug 2020).
- [20] Kaggle, “Diabetic retinopathy detection challenge,” (2015).
- [21] Decencière, E., Zhang, X., Cazuguel, G., Laÿ, B., Cochener, B., Trone, C., Gain, P., Ordóñez-Varela, J. R., Massin, P., Erginay, A., Charton, B., and Klein, J. C., “Feedback on a publicly distributed image database: The Messidor database,” *Image Analysis and Stereology* **33**, 231–234 (aug 2014).
- [22] Yan, Z., Han, X., Wang, C., Qiu, Y., Xiong, Z., and Cui, S., “Learning mutually local-global u-nets for high-resolution retinal lesion segmentation in fundus images,” *Proceedings - International Symposium on Biomedical Imaging* **2019-April**, 597–600 (2019).
- [23] Sarhan, M. H., Albarqouni, S., Yigitsoy, M., Navab, N., and Eslami, A., “Multi-scale microaneurysms segmentation using embedding triplet loss,” in [*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*], **11764 LNCS**, 174–182 (2019).
- [24] Guo, S., Li, T., Kang, H., Li, N., Zhang, Y., and Wang, K., “L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images,” *Neurocomputing* **349**, 52–63 (2019).
- [25] Guo, S., Li, T., Wang, K., Zhang, C., and Kang, H., “A lightweight neural network for hard exudate segmentation of fundus image,” in [*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*], **11729 LNCS**, 189–199, Springer International Publishing (2019).
- [26] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F., “Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research,” *Data* **3**, 25 (jul 2018).
- [27] Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollar, P., “Focal Loss for Dense Object Detection,” in [*Proceedings of the IEEE International Conference on Computer Vision*], (2017).
- [28] Tan, M. and Le, Q. V., “EfficientNetV2: Smaller models and faster training,” in [*International Conference on Machine Learning*], 10096–10106 (2021).
- [29] Chen, H. and Shrivastava, A., “Group Ensemble: Learning an Ensemble of ConvNets in a single ConvNet,” *arXiv preprint arXiv:2007.00649* (2020).
- [30] Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., Wu, T., Xiao, J., Wang, F., Yin, B., Wang, Y., Danala, G., He, L., Choi, Y. H., Lee, Y. C., Jung, S.-H., Li, Z., Sui, X., Wu, J., Li, X., Zhou, T., Toth, J., Baran, A., Kori, A., Chennamsetty, S. S., Safwan, M., Alex, V., Lyu, X., Cheng, L., Chu, Q., Li, P., Ji, X., Zhang, S., Shen, Y., Dai, L., Saha, O., Sathish, R., Melo, T., Araújo, T., Harangi, B., Sheng, B., Fang, R., Sheet, D., Hajdu, A., Zheng, Y., Mendonça, A. M., Zhang, S., Campilho, A., Zheng, B., Shen, D., Giancardo, L., Quelled, G., and Mériaudeau, F., “IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge,” *Medical Image Analysis* **59**, 101561 (jan 2020).

- [31] Chollet, F., “Xception: Deep learning with depthwise separable convolutions,” in [*Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*], (2017).
- [32] Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., and Kang, H., “Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening,” *Information Sciences* **501**, 511–522 (oct 2019).
- [33] Hu, J., Shen, L., and Sun, G., “Squeeze-and-excitation networks,” in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 7132–7141 (2018).
- [34] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L. C., “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in [*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*], 4510–4520 (2018).

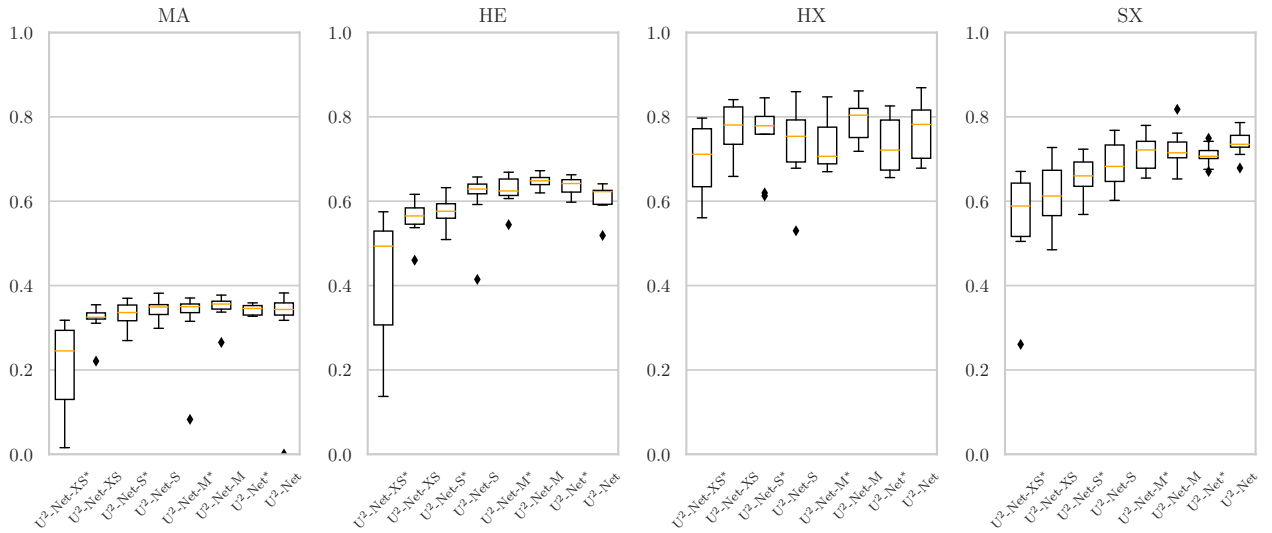


Figure 3. Boxplots of the AUPR score of the scaled U²-Nets over ten training runs for each lesion. Sorted by model size in ascending order from left to right.

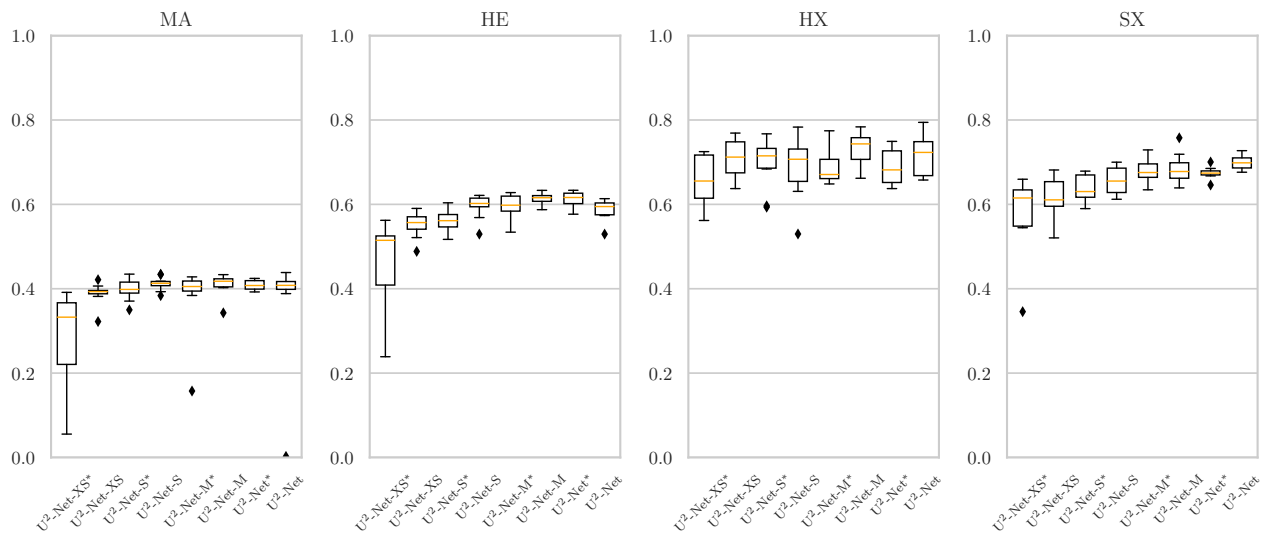


Figure 4. Boxplots of the Dice-score of the scaled U^2 -Nets over ten training runs for each lesion. Sorted by model size in ascending order from left to right.

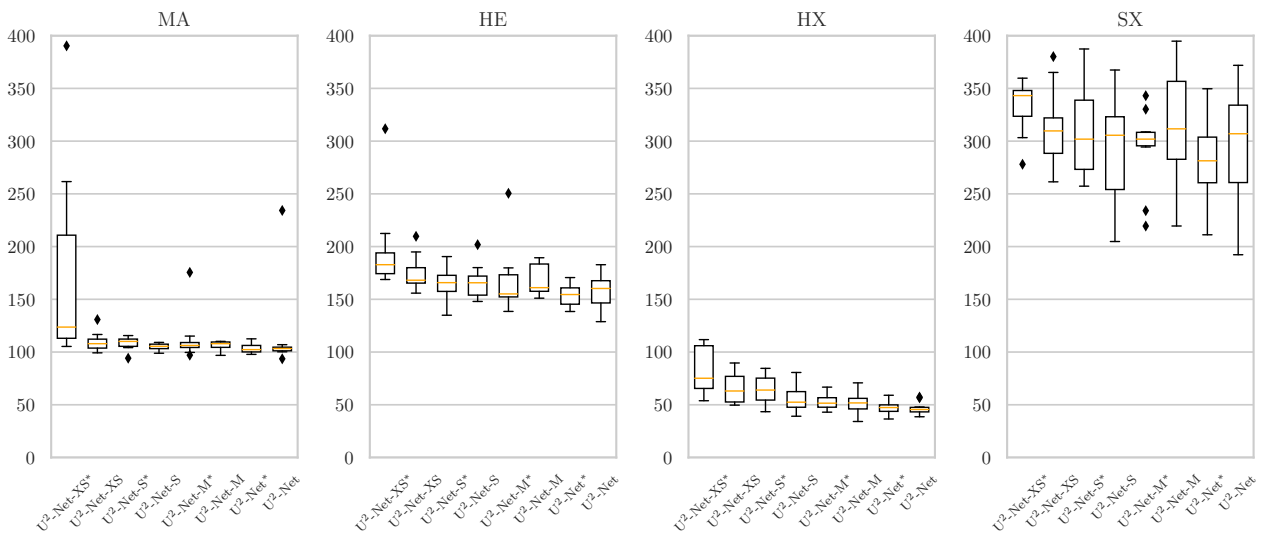


Figure 5. Boxplots of the Hausdorff-distance of the scaled U^2 -Nets over ten training runs for each lesion. Sorted by model size in ascending order from left to right.