Proceedings of Machine Learning Research 116:171–183, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection

Paul F. Jaeger[1] P.JAEGER@DKFZ.DE
Simon A. A. Kohl[1*] SIMONAAKOHL@GMAIL.COM
Sebastian Bickelhaupt[2] S.BICKELHAUPT@DKFZ.DE
Fabian Isensee[1] F.ISENSEE@DKFZ.DE
Tristan Anselm Kuder[3] T.KUDER@DKFZ.DE
Heinz-Peter Schlemmer[2] H.SCHLEMMER@DKFZ.DE
Klaus H. Maier-Hein[1] K.MAIER-HEIN@DKFZ.DE

[1] *Division of Medical Image Computing, German Cancer Research Center, Heidelberg, Germany*
[2] *Department of Radiology, German Cancer Research Center, Heidelberg, Germany*
[3] *Medical Physics in Radiology, German Cancer Research Center, Heidelberg, Germany*
[*] *Now with Karlsruhe Institute of Technology and DeepMind*

Editors: Adrian V. Dalca, Matthew Mcdermott, Emily Alsentzer, Sam Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

The task of localizing and categorizing objects in medical images often remains formulated as a semantic segmentation problem. This approach, however, only incompletely solves the object-level task, which then requires additional ad-hoc heuristics for mapping pixel- to object-level scores. State-of-the-art object detectors allow for individual object scoring in an end-to-end fashion. Ironically, being supervised by bounding box annotations, they currently achieve this only by trading in potentially precious pixel-wise annotations. This can be particularly disadvantageous in the setting of medical image analysis, where data sets are notoriously small and pixel-wise annotations are commonly available. In this paper, we propose *Retina U-Net*, a simple architecture, which naturally fuses the Retina Net one-stage detector with the U-Net architecture widely used for semantic segmentation in medical images. The proposed architecture combines object detection with an auxiliary semantic segmentation task, thus achieving end-to-end object-level analysis while leveraging full segmentation supervision. Our evaluation comprises in-depth comparisons to the most prevalent object detectors in 2D and 3D (code available at https://github.com/pfjaeger/medicaldetectiontoolkit) using two independent medical data sets (1035 lung CTs and 331 breast MRIs) and a series of toy experiments.

1. Introduction

Semantic segmentation algorithms, such as the U-Net architecture (Ronneberger et al. (2015)), are essential for tasks like radiation therapy planning or tumor growth monitoring, where pixel-wise information is clinically required. In many settings, however, clinicians are interested in "how many lesions were detected or missed and what is their grading score?"

rather than in information corresponding to individual pixels. In other words, in such scenarios downstream decisions are made on an object-level, requiring potentially brittle post-processing to bridge between pixel-wise predictions from semantic segmentation and object-level evaluation. To avoid this and allow end-to-end object scoring directly within one model, the state of the art converged to the natural solution of deriving object-level predictions from coarser representation levels of feature pyramid networks (FPNs), an architecture that most current object detectors are based on (Lin et al. (2017)). So-called two-stage detectors first discriminate objects from background while simultaneously regressing bounding box coordinates (He et al. (2017); Liu et al. (2018); Dai et al. (2016)). Subsequently proposals are categorized after resampling them to a fixed grid, thus ensuring scale-invariance for categorization. One-stage detectors, on the other hand, have been proposed to perform categorization directly on the coarse FPN representations, in concurrence to the box regression (Liu et al. (2016); Lin et al. (2018); Redmon et al. (2016)). Both approaches come at the price of disregarding potentially helpful pixel-wise annotations, which are reduced to bounding boxes (or cubes). This information loss contradicts the need for data-efficient training especially in the medical domain, where comparably small data sets are quite common. Previous work in the non-medical domain has addressed this problem (Shrivastava and Gupta (2016); Mao et al. (2017); Dvornik et al. (2017)), but has either used the extra supervision signal in a sub-optimal fashion (e.g. at lower resolutions or without surrounding context) or has employed approaches of arguably significant model complexity.

This study demonstrates a straight-forward yet effective strategy on how to convert available pixel-wise annotations into significant performance gains on object-level detection tasks. Retina U-Net is based on the plain one-stage detector Retina Net, which is complemented by architectural elements of the U-Net, a very successful model for semantic segmentation of medical images (Ronneberger et al. (2015)). Specifically, the decoding part of Retina Net is complemented by additional high resolution levels to learn the auxiliary task of semantic segmentation. From a segmentation perspective, the proposed architecture retrofits the U-Net with two sub-networks operating on coarser feature levels of the decoder part to allow for end-to-end object scoring. Regarding the choice of a one-stage detector, we argue that the explicit scale invariance enforced by the resampling operation in two-stage detectors is not helpful in the medical domain, since unlike in natural images scale is not a function of varying distances between object and camera, but encodes semantic information. We demonstrate the effectiveness of our model on the task of detecting and categorizing lesions on two medical data sets and support our analysis by a series of toy experiments that help shed light on the reasons behind the observed performance gains. The contributions are the following:

- A simple but effective method for leveraging semantic segmentation training signals in object detection focused on application in medical images.
- An in-depth analysis of the prevalent object detectors (operating in 2D as well as 3D) by means of comparative studies on medical data sets.
- A comprehensive framework including e.g. modular implementations of all explored models, with code available at
  https://github.com/pfjaeger/medicaldetectiontoolkit.

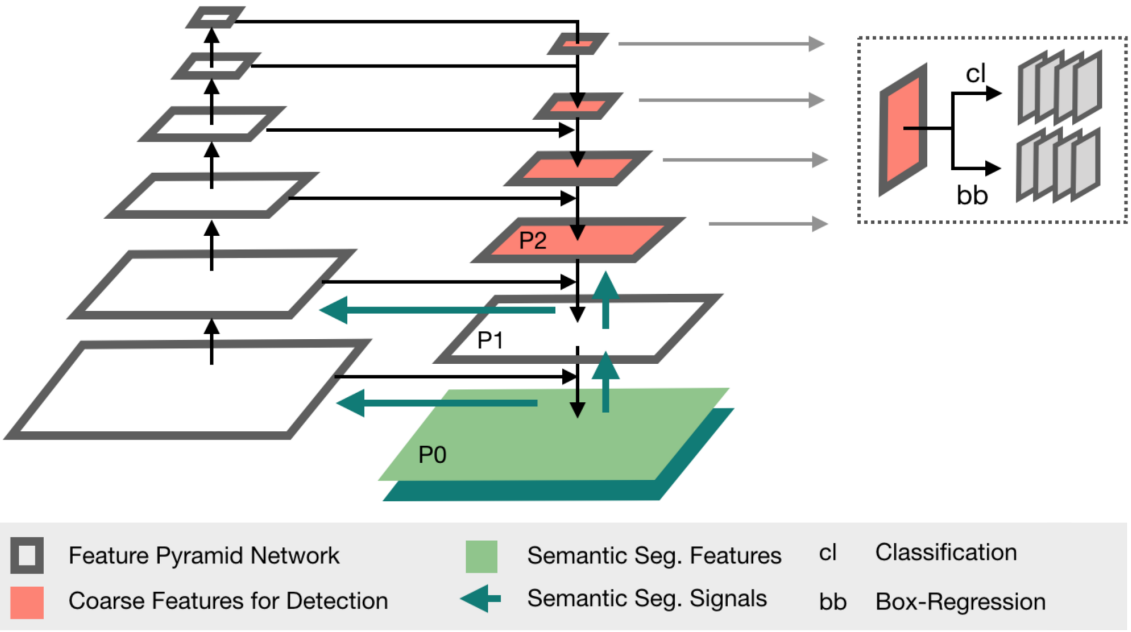

Figure 1: The Retina U-Net architecture in 2D. From a Retina Net perspective we add pyramid levels (i.e. feature maps) P0 and P1, so as to backpropagate rich training signals from a full resolution pixel-wise loss into the feature pyramid network, thereby facilitating the learning process in coarser feature maps used for object detection. From a U-Net perspective we enable end-to-end object detection via a head network operating on the coarse pyramid levels P2-P5.

2. Related Work

Since object detection in natural images is increasingly formulated as an instance segmentation problem, several two-stage object detectors utilize additional instance-based segmentation labels during training (He et al. (2017); Chen et al. (2018a); Liu et al. (2018)). However, we argue that this setup does not fully exploit semantic segmentation supervision:

- The mask loss is only evaluated on cropped proposal regions, i.e. context gradients of surrounding regions are not backpropagated.
- The proposal region as well as the ground truth mask are typically resampled to a fixed-sized grid (known as RoIAlign (He et al. (2017))).
- Only positive matched proposals are utilized for the mask loss, which induces a dependency on the region proposition performance.
- Gradients of the mask loss do not flow through the entire model, but merely from the corresponding pyramid level upwards.

Auxiliary tasks for exploiting semantic segmentation supervision have been applied in two stage detectors with bottom-up feature extractors (i.e. encoders) (Shrivastava and Gupta (2016); Mao et al. (2017)). In the one-stage domain, the work of Uhrig et al. (2018) performs semantic segmentation on top of a single-shot detection (SSD) architecture for instance segmentation, where segmentation outputs are assigned to box proposals in a post-processing step. Zhang et al. (2018) propose a similar architecture, but learn segmentation in a weakly-supervised manner, using pseudo-masks created from bounding box annotations.

As opposed to bottom-up backbones for feature extraction, we follow the argumentation of feature pyramid networks (Lin et al. (2017)), where a top-down (i.e. decoder) pathway is installed to allow for semantically rich representations at different scales. This concept is adapted from state-of-the-art segmentation architectures (Ronneberger et al. (2015); Chen et al. (2018b)) and used in both current one- and two-stage detectors. Recent approaches report to use semantic segmentation as an auxiliary training signal, but apply it to lower resolution feature maps only (Peng et al. (2018); Dvornik et al. (2017)). Araújo et al. (2018) learn semantic segmentation at full resolution, but downsample all feature maps to the bottleneck's spatial resolution before feeding them to the detection module, thereby discarding all multi-scale information in the detection task. In contrast, we propose a FPN-based one-stage detector performing semantic segmentation on full resolution and object detection on multiple scales, which allows to naturally fuse existing state-of-the-art models from both domains resulting in the simple Retina U-Net architecture. Shah et al. (2018) converged to an architecture similar to ours while working on the inverse setup, i.e. enhancing segmentation performance by additionally training on bounding box labels.

3. Methods

3.1. Retina Net

The basis of our proposed model is the Retina Net, a simple one-stage detector based on a FPN for feature extraction (Lin et al. (2018)), where two sub-networks operate on the pyramid levels P_3-P_6 for classification and bounding box regression, respectively (see Figure 2c). Here P_j denotes the feature-maps of the jth decoder level, where j increases as the resolution decreases. In this study, we compare various state-of-the-art one- and two-stage object detectors in both 2D (slice-based) and 3D (volumetric patches). For the sake of unrestricted comparability, all methods including Retina Net are implemented in one framework, using the same backbone FPN (Lin et al. (2017)) based on a ResNet50 (He et al. (2016)) as identical architecture for feature extraction. In our FPN implementation, anchor sizes are divided by a factor of 4 to account for smaller objects in the medical domain resulting in anchors of size $\{4^2, 8^2, 16^2, 32^2\}$ for the corresponding pyramid levels $\{P_2, P_3, P_4, P_5\}$. In the 3D implementation, the z-scale of anchor-cubes is set to $\{1, 2, 4, 8\}$. For Retina Net, two adaptations were made deviating from the original version: To factor in the existence of small object sizes in medical images, we shifted sub-network operations by one pyramid level towards P_2-P_5. This comes at a computational price, since a vast number of dense positions are produced in the higher resolution P_2 level. We further exchanged the sigmoid non-linearity in the classification sub-network for a softmax operation, to account for mutual exclusiveness of classes due to non-overlapping objects in 3D images.

3.2. Retina U-Net

Retina Net is complemented with architectural elements from the U-Net, resulting in the proposed Retina U-Net. Training signals for full semantic supervision are added to the top-down path by means of additional pyramid levels P_1 and P_0, including the respective skip connections. The resulting Feature Pyramid resembles the symmetric U-Net architecture (see Figure 1), which in the following we refer to as *U-FPN* for clarity. The detection sub-networks do not operate on P_1 and P_0, which keeps the number of parameters at inference time unchanged. The segmentation loss is calculated from P_0 logits. In addition to a pixel-wise cross entropy loss $\mathcal{L}_{\text{CE}}$, a soft Dice loss is applied, which has been shown to stabilize training on highly class imbalanced segmentation tasks e.g. in the medical domain (Isensee et al. (2018)):

$$\mathcal{L} = \mathcal{L}_{\text{CE}} - \frac{2}{|K|} \sum_{k \in K} \frac{\sum_{i \in I} u_{ik} v_{ik}}{\sum_{i \in I} u_{ik} + \sum_{i \in I} v_{ik}}, \tag{1}$$

where u is the softmax output of the network and v is a one hot encoding of the ground truth segmentation map. Both u and v have shape $I \times K$ with $i \in I$ being the number of pixels in the training batch and $k \in K$ being the classes.

3.3. Baseline methods.

The following baseline methods were implemented:

- Mask R-CNN (He et al. (2017)): Adjustments for the 3D implementation: The number of feature maps in the region proposal network is lowered to 64 to account for increased GPU memory usage. The poolsize of 3D-RoIAlign, a 3D implementation of the resampling operation mentioned in Section 1, is set to (7, 7, 3) for the classification head and (14, 14, 5) for the mask head. The matching IoU for positive proposals is lowered to 0.3 (see Figure 2a).

- Faster R-CNN+: In order to single out the Mask R-CNN's performance gain obtained by segmentation supervision from the mask head, we run ablations on the toy data sets while disabling the mask-loss, thereby effectively reducing the model to the Faster R-CNN architecture (Girshick (2015)) except for the RoIAlign operation (indicated in the method's name by an additional +) (see Figure 2b).

- Retina U-Net$_{\text{2stage}}$: To analyze the gain of additional semantic segmentation supervision in two-stage detectors, we implemented a two-stage variant of Retina U-Net by deploying Faster R-CNN+ on top of U-FPN. (see Figure 2d).

- U-Net+Heuristics: Essentially formulating the problem as a semantic segmentation task, as commonly done in medical imaging, we implement a U-Net-like baseline using

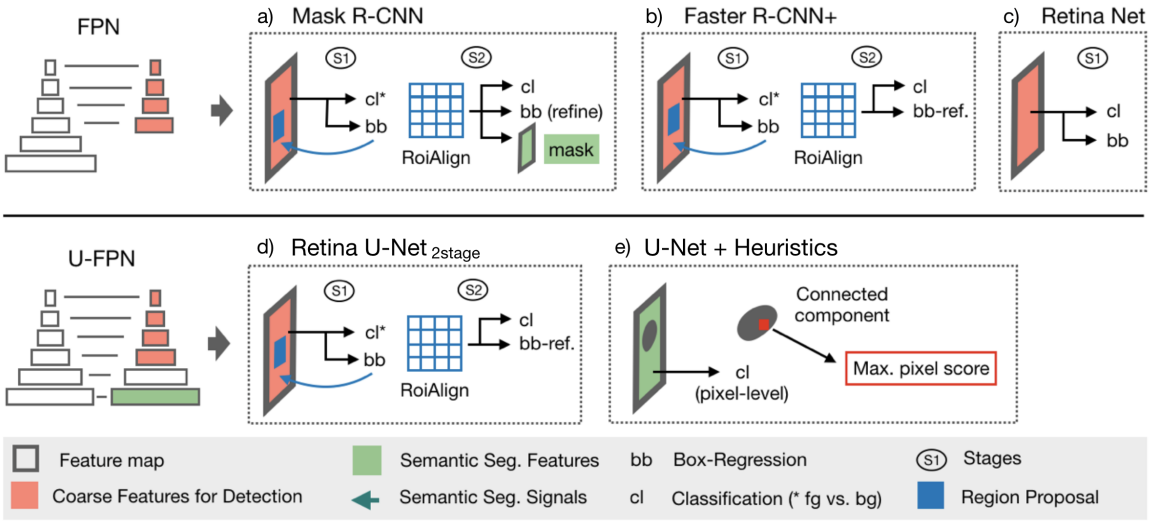

Figure 2: The upper panel shows all baselines utilizing a regular FPN feature extractor while the lower panel depicts baselines that employ a symmetric FPN akin to a U-Net (U-FPN). Subfigures a) - e) show the detection sub networks (heads) that are characteristic of each model and operate on FPN features. All models employ their respective head topology to different decoder scales which are denoted in red. Boxes in green indicate logits that are trained on an auxiliary semantic segmentation task.

U-FPN. Therefore, softmax predictions were extracted from P_0 via 1x1 convolution and utilized to identify connected components for all foreground classes. Subsequently, bounding boxes (or cubes) are drawn around connected components and the highest softmax probability per component and class is assigned as object score (see Figure 2e).

4. Experiments

4.1. Clinical Datasets

A lung nodule detection and categorization task is performed on the publicly available LIDC-IDRI data set (Armato III et al. (2015)), consisting of 1035 lung CT scans with pixel-wise lesion annotations and malignancy likelihood scores (1-5) from four experts. We fuse annotations of raters per lesion by applying a pixel-wise majority voting and averaging the malignancy scores. Scores are then re-labelled into benign (1-2: 1319 cases) and malignant (3-5: 494 cases). A breast lesion detection and categorization task is performed on an in-house Diffusion MRI data set of 331 Patients with suspicious findings in previous mammography. Pixel-wise annotations of lesions are provided by experts. Categorisation labels are given by subsequent pathology results (benign: 141 cases, malignant: 190 cases).

Both clinical data sets require the detection and categorisation (*benign* vs. *malignant*) of lesions. This is both a difficult and very frequent problem setting in radiology and therefore constitutes a highly-relevant domain of application. Example images with model predictions are shown in Figure 3

4.2. Toy Datasets

We created a series of three toy experiments to separately evaluate on sub-tasks commonly involved in object-categorisation on medical images (see Figure 4). More specifically, the aim is to investigate the importance of full segmentation supervision in the context of limited training data:

1. *Distinguishing object shapes:* Two classes of objects are to be detected and distinguished, circles and donuts (cut-out hole in the middle). Here, the corresponding segmentation mask's shape explicitly contains the discriminative feature (the cut-out hole), hence, full semantic supervision is expected to yield significant performance gains.

2. *Learning discriminative object patterns:* This task is identical to the previous one, except the central hole is not cut out from the segmentation masks of the donuts (class 2). This requires the model to pick up the discriminative pattern without explicitly receiving the respective training signal by means of the mask's shape. This setup could be considered more realistic in the context of medical images.

3. *Distinguishing object scales:* Circles of two different sizes (19 vs. 20 pixel diameter) are to be detected and distinguished. Here, class information is entirely encoded in object scales and hence in target box coordinates. No significant gain from semantic supervision is expected.

Each toy data set consists of 2500 artificially generated 2D images of size 320×320 (1000 train / 500 val / 1000 hold-out test). Images are zero-initialized and foreground objects imprinted by increasing intensity values by 0.2. Subsequently, uniform noise is added to all pixels.

4.3. Training & Evaluation Setup.

For comparability, experiments for all methods are run with identical training and inference schemes, as described below. In this study, we compare slice-wise 2D processing, feeding the ± 3 neighbouring slices as additional input channels (2Dc) (Yan et al. (2018)), against 3D convolutions. Oversampling of foreground regions is applied when training on patch crops. To account for the class-imbalance of object level classification losses, we stochastically mine the hardest negative object candidates according to softmax probability. Models are trained in a 5-fold cross validation (splits: train 60% / val 20% / test 20%) with batch size 20 (8) in 2D (3D) using the Adam optimizer (with default settings) at a learning rate of 10^{-4}. Extensive data augmentation in 2D and 3D is applied to account for overfitting. To compensate for unstable statistics on small data sets, we report results on the aggregated inner loop test sets and ensemble by performing test-time mirroring as well as by testing on multiple models selected as the 5 highest scoring epochs according to validation metrics.

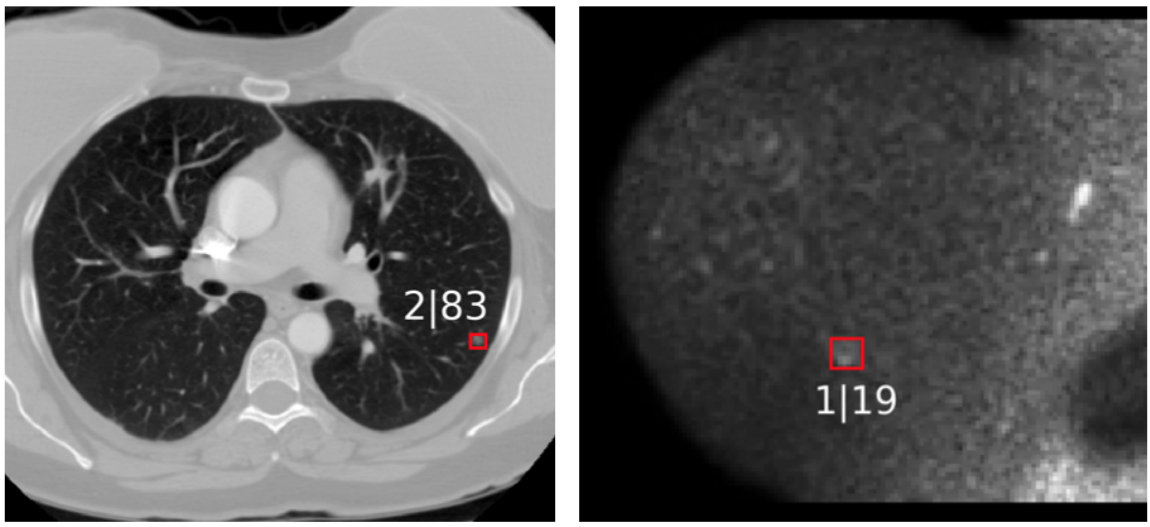

Figure 3: Example predictions of a malignant lesion in a CT scan of the lung (left) and a benign lesion on a Diffusion MRI of the breast (right). The numbers before and after the vertical bar denote the predicted class-id and the prediction confidence in percent, respectively.

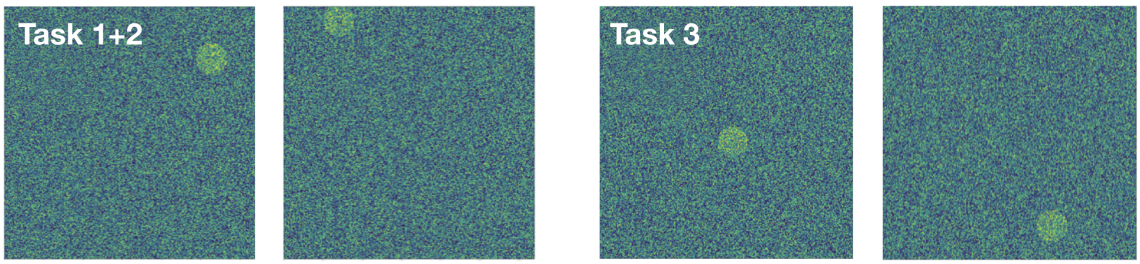

Figure 4: *left:* Example images for tasks 1 (distinguishing object shapes) and 2 (learning discriminative object patterns) of the toy experiment series. The left object is a filled circle, while the object on the right is a donut (cut-out hole in the middle). *right:* Example images for task 3 (distinguishing object scales), the circle on the left has a diameter of 19 pixels, as opposed to 20 pixels diameter of the circle on the right.

Consolidation of box predictions from ensemble-members and overlapping tiles is done via clustering and weighted averaging of scores and coordinates. Experiments are evaluated using mean average precision (mAP) (Everingham et al. (2010)). We determine mAP at a relatively low matching intersection over union (IoU) threshold of IoU = 0.1, which

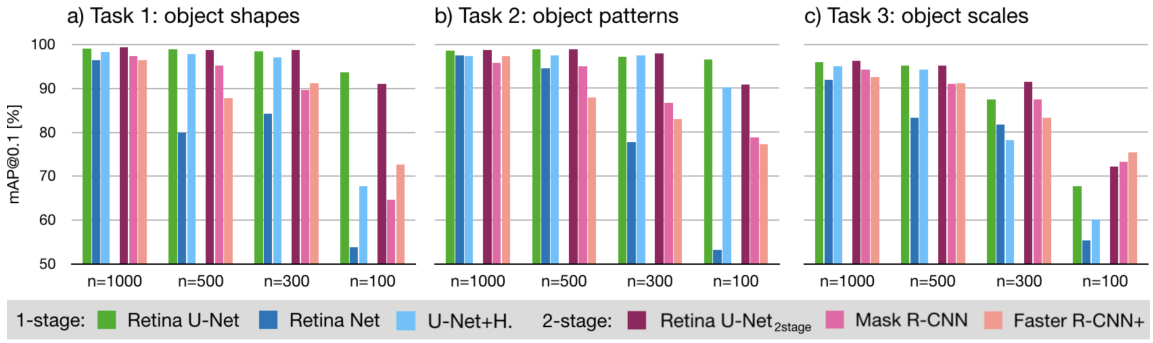

Figure 5: Results of the of toy experiment series. The three tasks are displayed as (a) distinguishing objects of different shapes, (b) learning discriminative image patterns unrelated to an object's shape, and (c) distinguishing objects of different scales. Explored models are divided into two groups: One-stage methods have blue/green color, while two-stage methods are drawn in red. n denotes the number of training images.

Table 1: Results for lung lesion detection on CT (LIDC) and breast lesion detection on DWI, reported as [%].

		LIDC		Breast DWI	
Dim.	Model	mAP_{10}	AP_{pat_m}	mAP_{10}	AP_{pat_m}
2Dc	Retina U-Net (ours)	**50.2**	**73.9**	**33.4**	**86.9**
	Mask R-CNN	45.4	69.1	32.3	86.4
	Retina Net	48.2	71.5	33.2	84.4
	Retina U-Net$_{2stage}$ (ours)	49.1	71.6	33.2	84.7
	U-Net+Heuristics	41.1	66.1	25.8	81.6
3D	Retina U-Net (ours)	49.8	70.4	**35.8**	**88.0**
	Mask R-CNN	48.3	**71.8**	34.0	84.8
	Retina Net	45.9	68.8	31.9	86.4
	Retina U-Net$_{2stage}$ (ours)	**50.5**	70.7	35.1	86.5
	U-Net+Heuristics	36.6	62.8	26.9	85.1

respects the clinical need for coarse localization and also exploits the non-overlapping nature of objects in 3D. Note that evaluation and matching is performed in 3D for all models and processing setups. Additionally, we report the AP of patient-scores, which are determined as the maximum of scores per class and patient (aggregating predictions as well as labels over potentially multiple lesions per patient). In our setting, however, patient scoring is to be taken with a grain of salt: Since the information about assignment of overall-patient

scores to the specific lesion is lost in the aggregation process, this metric is blind to the issue of 'being right for the wrong reasons' and further potentially biased by class-imbalances.

5. Results & Discussion

5.1. Clinical Datasets

Results for the lesion detection and categorization tasks are shown in Table 1. Retina U-Net performs best on the 2Dc setups (50.2 mAP on LIDC and 33.4 mAP on Breast DWI). In 3D, Retina U-Net performs best on Breast DWI (35.8 mAP) and only slightly worse (49.8 mAP) than Retina U-Net$_{2stage}$ (50.5 mAP) on LIDC. Comparing these results to the remaining baselines shows the importance of semantic segmentation supervision. Mask-R-CNN for instance, which is trained with instance segmentation supervision instead, shows overall worse performance presumably due to the issues discussed in Section 2. U-Net performs worse with notable margins, seemingly suffering from high confidence false positive predictions caused by the necessary max-score aggregation (aggregating scores via mean in an alternative experiment did not improve performance). Evaluating on patient-level, Retina U-Net performs best on all tasks except 3D LIDC, where Mask R-CNN achieves the highest score. This finding could be related to the brittle aggregation heuristics of patient scores described above.

5.2. Toy Datasets

Results for the toy experiments are shown in Figure 5. In the first task, where explicit class information is contained in segmentation annotations, models which optimally leverage those, i.e. Retina U-Net and Retina U-Net$_{2stage}$, perform best (again, the instance-based segmentation training of Mask R-CNN seems to yield inferior performance presumably due to issues discussed in Section 2). The resulting margin increases with decreasing amount of available training data. The second task, where class information is effectively removed from segmentation annotations, shows similar margins of Retina U-Net and Retina U-Net$_{2stage}$ to other models. This indicates the importance of full semantic segmentation supervision even in implicit setups and shows a particularly strong edge in the small data set regime, where models that discard this supervision essentially collapse. In the third task, where class information is entirely contained in the target boxes, no gain from segmentation supervision is observed, at least for small training data sets. Surprisingly, two-shot detectors perform better at this task, which seems counter-intuitive given the scale-invariance enforced by the RoIAlign operation. We hypothesize, that discarding the spatial scale effectively forces the optimizer to encode this information into the feature maps previous to the RoIAlign operation, which seems to not hamper performance and possibly even cause a beneficial regularization effect. Comparing Mask R-CNN to Faster R-CNN+, the sub-optimal mask-supervision seems to yield no gains in detection performance when working with limited training data.

6. Conclusion

In this paper, we address the challenge of aligning the output structure of predictive models to the requirements of the intended clinical task while maintaining data efficient training on small data sets. Specifically, we propose Retina U-Net, a simple but effective method for leveraging segmentation supervision in object detection. We show the importance of exploiting these training signals on multiple data sets, input dimensions and meticulously compare against the prevalent object detection models with a particular emphasis on the context of limited training data. Therefore, we consider the task of localizing and classifying lesions, which constitutes a difficult and very frequent problem setting in radiology and therefore a highly-relevant domain of application. On the publicly available LIDC-IDRI lung CT dataset as well as on our in-house breast lesion MRI dataset, Retina U-Net yields detection performance superior to models without full segmentation supervision. By means of a set of toy experiments we shed light on an important set of scenarios that can profit from the additional full supervision: Any such problem where there is discriminative power in features beyond mere scale can expect to pocket an edge in detection performance. Among other distinguishing characteristics, the domain of medical image analysis holds one prominent feature: scarcity of labelled data. Retina U-Net is designed to make the most of the given supervision signal which is a key advantage on small datasets as high-lighted by our experiments. Our architecture stands out with another feature: its embarrassingly simple model formulation that takes aim at broad applicability and lays the foundation for future work on clinical requirements such as interpretability and robustness.

References

Teresa Araújo, Guilherme Aresta, Adrian Galdran, Pedro Costa, Ana Maria Mendonça, and Aurélio Campilho. Uolo-automatic object detection and segmentation in biomedical images. In *DLMIA*, pages 165–173. Springer, 2018.

SG Armato III, G McLennan, L Bidaut, MF McNitt-Gray, CR Meyer, AP Reeves, and LP Clarke. Data from lidc-idri. the cancer imaging archive. *DOI http://doi. org/10.7937 K*, 9, 2015.

Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, June 2018a.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *ECCV*, 2018b.

Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016.

Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. Blitznet: A real-time deep network for scene understanding. In *ICCV*, Oct 2017.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, Jun 2010. ISSN 1573-1405.

Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988. IEEE, 2017.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *TPAMI*, 2018.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

Jiayuan Mao, Tete Xiao, Yuning Jiang, and Zhimin Cao. What can help pedestrian detection? In *CVPR*, pages 6034–6043. IEEE, 2017.

Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *CVPR*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

Meet P. Shah, S. N. Merchant, and Suyash P. Awate. Ms-net: Mixed-supervision fully-convolutional networks for full-resolution segmentation. In *MICCAI*, pages 379–387. Springer, 2018.

Abhinav Shrivastava and Abhinav Gupta. Contextual priming and feedback for faster r-cnn. In *ECCV*, pages 330–348. Springer, 2016.

Jonas Uhrig, Eike Rehder, Björn Fröhlich, Uwe Franke, and Thomas Brox. Box2pix: Single-shot instance segmentation by assigning pixels to object boxes. In *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

Ke Yan, Mohammadhadi Bagheri, and Ronald M Summers. 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In *MICCAI*, pages 511–519. Springer, 2018.

Zhishuai Zhang, Siyuan Qiao, Cihang Xie, Wei Shen, Bo Wang, and Alan L Yuille. Single-shot object detection with enriched semantics. Technical report, CBMM, 2018.