2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

SYNERGYNET: A FUSION FRAMEWORK FOR MULTIPLE SCLEROSIS BRAIN MRI SEGMENTATION WITH LOCAL REFINEMENT

Yeeleng S. Vang[⋆ †] *Yingxin Cao*[⋆ †] *Peter D. Chang*[‡] *Daniel S. Chow*[‡]
Alexander U. Brandt[¶] *Friedemann Paul*[¶] *Michael Scheel*[¶] *Xiaohui Xie*[†]

[†] Dept of Computer Science, University of California, Irvine CA 92697
[‡] University of California, Irvine Medical Center, 101 The City Drive S. Bldg. 55 201, Orange, CA 92868
[¶] NeuroCure Clinical Research Center, Charité –Universitätsmedizin Berlin corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Germany

ABSTRACT

The high irregularity of multiple sclerosis (MS) lesions in sizes and numbers often proves difficult for automated systems on the task of MS lesion segmentation. Current State-of-the-art MS segmentation algorithms employ either only global perspective or just patch-based local perspective segmentation approaches. Although global image segmentation can obtain good segmentation for medium to large lesions, its performance on smaller lesions lags behind. On the other hand, patch-based local segmentation disregards spatial information of the brain. In this work, we propose SynergyNet, a network segmenting MS lesions by fusing data from both global and local perspectives to improve segmentation across different lesion sizes. We achieve global segmentation by leveraging the U-Net architecture and implement the local segmentation by augmenting U-Net with the Mask R-CNN framework. The sharing of lower layers between these two branches benefits end-to-end training and proves advantages over simple ensemble of the two frameworks. We evaluated our method on two separate datasets containing 765 and 21 volumes respectively. Our proposed method can improve 2.55% and 5.0% for Dice score and lesion true positive rates respectively while reducing over 20% in false positive rates in the first dataset, and improve in average 10% and 32% for Dice score and lesion true positive rates in the second dataset. Results suggest that our framework for fusing local and global perspectives is beneficial for segmentation of lesions with heterogeneous sizes.

Index Terms— Multiple Sclerosis, Deep Learning, MRI

1. INTRODUCTION

Multiple Sclerosis (MS) is a prevalent chronic autoimmune disease affecting the central nervous system (CNS), characterized by white matter or gray matter lesions formed during inflammation and demyelination process of the brain and spinal cord [1]. MS lesions can be detected using magnetic resonance imaging (MRI). Most lesions display hyper-intense appearance under T_2-w and fluid attenuated inversion recovery (FLAIR) MR images, which makes MRI a standard diagnostic tool of MS [2]. Previous studies show that determination of volume and spatial location of lesions is important for diagnosing and tracking of the disease. Although being considered as the gold standard for lesion segmentation, manual delineation of 3D medical images is time-consuming, labor-intensive and subjective sometimes [3, 4]. To this end, many automated segmentation methods have also been developed [5, 6, 7, 8] with deep learning-based methods heavily favored among the state-of-the-art systems [9, 10, 11, 12, 13, 14, 15].

In [10], a cascading, 3D patch-based CNN architecture is proposed with the first network generating initial probabilistic prediction and second network performing false positive reduction. In [9], a fully convolutional neural network is proposed using multi-channel 2D patches as input and outputs are concatenated feature maps producing lesion segmentation. While obtaining good segmentation performances, these patch-based local approaches never utilize the inherent spatial information from the entire brain. In constrast, in [11], convolutional restricted Boltzmann machines is used to pre-train the encoder branch of a U-Net style network. However, they do not take advantage of the middle layers which have been shown significant to segmentation performance [16]. In [12], a U-Net style network is used with three encoding branches corresponding to three different modality inputs followed by multi-scale feature upsampling and concatenation along the decoding patch. These global view approaches constitute the state-of-the-art systems, however they lack explicit mechanism to deal with problematic, smaller lesions.

In this paper, we propose a new model to tackle the problem of MS lesion segmentation in 3D FLAIR MRIs. We use U-Net as the global perspective segmentator. For smaller lesions, we augment the U-Net network with the Mask R-CNN framework as a mechanism for localized attention. We then fuse the global image segmentation outputted from U-

⋆ equal contribution

Net branch and the local patch segmentation outputted from Mask R-CNN branch via weighted averaging to achieve a final whole MR image segmentation. Fusing the two models under our framework is beneficial as both branches are jointly trained in an end-to-end fashion which is shown to be advantageous over simple ensemble of the two separate models.

Our main contributions are three folds. First, we proposed a novel method, which we call SynergyNet, tackling MS lesion irregularity in automated segmentation tasks, especially for small lessons where previous global segmentation methods are problematic. Second, we show how to achieve global and local data fusion by fusing U-Net and Mask R-CNN under a unified framework for end-to-end learning and achieve the state-of-the-art result for MS lesion segmentation. Third, this is the first relatively large scale MS segmentation study consisting of several hundred volumes for use in comparing deep learning frameworks.

2. PROPOSED METHOD

We propose a fusion network we call SynergyNet as a framework for 3D MS MRI lesion segmentation that maintains the good performance of U-Net for medium to large-size lesions and augments it with Mask R-CNN to improve the segmentation performance on small lesions.

2.1. SynergyNet

For global image segmentation, we choose U-Net style network as the backbone. Standard vanilla U-Net model performs well when segmenting medium-size to large, contiguous MS lesions, however they struggle with small lesions given the class imbalance, see Figure 2. Specifically with MS pathology, lesions can be as small as 10 voxels in a MRI containing as few as a handful or as many as 100s of such small lesions in addition to possibly containing much larger lesions. We replace the standard convolutional layer to residual blocks that has been proven effective [17].

To improve the segmentation performance of small lesions, we propose augmenting the U-Net backbone with an auxiliary network that focuses on detecting small lesions by using the Mask R-CNN framework as the auxiliary network. This framework works as a specialized detector and segmentator of specifically-sized lesions. Mask R-CNN have demonstrated good performance detecting and segmenting objects with various sizes provided these objects have more standardized shapes and aspect ratios such as in the natural image setting [18]. In contrast, MS lesions demonstrates a high degree of irregularity that makes it ill-posed for segmentation by Mask R-CNN. By restricting our focus to small lesions, we can effectively reduce the degree of irregularity in the subset of MS lesions we are interested in which minimizes Mask R-CNN's main draw back due to its use of a fixed set of rectangular-shaped anchor boxes. Here we define

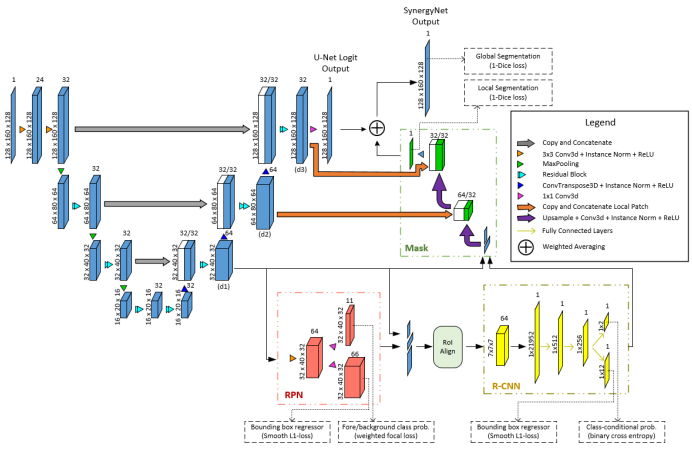

Fig. 1: Illustration of our SynergyNet framework. U-Net is employed as the global image segmentation backbone and Mask R-CNN is used to augment U-Net to provide localized attention for improved detection and segmentation of small lesions. The entire network is trained in an end-to-end fashion using a sequential additive training procedure.

small lesions as those with ground truth bounding box volume less than 1500 voxels. This specialized, local-focused branch works by running the RPN sub-network over the (d1) feature map of the decoding branch of U-Net and proposes local patches with lesions. Positive predicted proposals are then passed to the R-CNN sub-network to refine bounding boxes estimations. Finally the refined proposals are used to obtained multi-scale versions of these local patches from the d1, d2, and d3 feature map layers for multi-scale data fusion and segmentation of these local patches in the Mask sub-network. See Figure 1 for reference to these feature maps layers.

2.2. Global-Local Data Fusion

We extract their corresponding patches from multiple scales of the U-Net decoder branch before upsizing and combine them to leverage information from different receptive field sizes. We fuse the predictions from the global and local perspective branches by ensembling the local patch predictions with their corresponding location in the global image via weighted averaging. Concretely, given local output $Y_m(i)$ from the Mask sub-network and the corresponding patch $Y_u(i)$ in the U-Net global logit map, we ensemble and update the $Y_u(i)$ patch as follows:

$$Y_u(i) = \gamma * Y_m(i) + (1 - \gamma) * Y_u(i) \quad (1)$$

where γ is a hyperparameter weighing each predictors. For our experiment, we set γ=0.3 based on the validation set results. SynergyNet's loss function is a multi-tasks loss function consisting of losses from each of the individual sub-networks:

$$L_{total} = L_{RPN} + L_{R\text{-}CNN} + L_{MASK} + L_{U\text{-}Net} \quad (2)$$

where losses of the individual sub-networks are defined as:

$$L_{RPN} = \alpha_1 \sum_i -r_i(1-\hat{r}_i)^{\beta}\log(\hat{r}_i) + \alpha_2\left(\frac{1}{J}\sum_j z_j\right)$$

$$L_{R\text{-}CNN} = -\alpha_3 \sum_c s_c \log(\hat{s}_c) + \alpha_4\left(\frac{1}{D}\sum_d z_d\right)$$

$$L_{MASK} = \alpha_5\left(\frac{1}{M}\sum_m 1 - \frac{2at}{a^2+t^2}\right)$$

$$L_{U\text{-}Net} = \alpha_6\left(\frac{1}{U}\sum_u 1 - \frac{2OT}{O^2+T^2}\right)$$

where $z_k = \begin{cases} 0.5(e_k-\hat{e}_k)^2, & \text{if } |e_k-\hat{e}_k| < 1 \\ |e_k-\hat{e}_k| - 0.5, & \text{otherwise} \end{cases}$ is the smooth L1 loss for $k = j$ or d. L_{RPN} consists of the focal loss [19] for binary classification and smooth L1 regression losses of the RPN network where r_i and $\hat{r}_i$ are the true and predicted probability of proposal i containing a lesion respectively, and e_j and $\hat{e}_j$ are the true and predicted delta coordinates of the bounding box for those proposals j predicted to contain lesions. β was set to 2 as in the original paper for focal loss [19]. L_{R-CNN} consists of the binary cross-entropy loss and smooth L1 regression losses of the class-conditional R-CNN sub-network where s_c and $\hat{s}_c$ are the true and predicted probability of proposal c outputted from RPN containing lesion respectively, and e_d and $\hat{e}_d$ are the true and predicted delta coordinates of the bounding volume for those proposals predicted to contain lesions. L_{MASK} consists of the 1-Dice loss of the Mask sub-network where a and t are the ground truth and predicted segmentation volume of the patch m respectively. L_{U-Net} consists of the 1-Dice loss of the post-ensembled SynergyNet output where O and T are the ground truth and predicted segmentation of the entire MR volume respectively. For this experiment, the following hyperparameters are used: $\alpha_1 = \alpha_2 = \alpha_4 = \alpha_5 = 1$, $\alpha_3 = 1.3$, and $\alpha_6 = 5$.

2.3. Training Procedure

During training, each sub-network is trained in a sequentially additive fashion until the entire SynergyNet is trained. More specifically, the RPN sub-network is train alone with the feature extraction network for the first 30 epochs, follow by the R-CNN sub-network in addition to the first two sub-networks for the next 30 epochs, then by including the Mask sub-network training for the following 40 epochs, and finally the remaining layers of the decoding branch of the U-Net sub-network added for the last 100 epochs. This is necessary to avoid the computational cost that would otherwise be required by the entire model during the first few training epochs where a large set of false positive proposals is made until they are learned to be rejected.

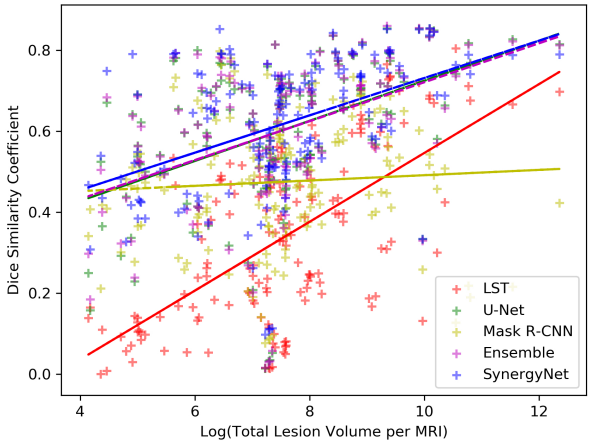

Fig. 2: Private test set dice score plotted as a function of the logarithm of total lesion volume per MRI. Trendlines are fitted to the scatter plot of each algorithms.

3. EXPERIMENTS AND RESULTS

Two separate datasets are used to evaluate our proposed model. The private dataset was obtained at Charité - Universitätsmedizin Berlin, Germany, from ongoing observational cohort studies of patients with autoimmune neuroinflammatory disorders. After cleansing for patients lack of ground truth mask, we are left with 765 longitudinal data points belonging to 261 patients with full 182x218x182 3D FLAIR and T_2-w MR images along with their ground truth lesion mask. We randomly partition the patient set into 60%-20%-20% training-validation-test resulting in 444 volumes for training, 166 volumes for validation, and 155 volumes for test. Original FLAIR MRIs are skull-stripped, bias corrected and standardized with z-score standardization. The public dataset consists of 21 longitudinal volumes from 5 unique patients as part of the 2015 ISBI longitudinal MS lesion segmentation challenge [4].

Five metrics are used to evaluate the segmentation performance of these five algorithms. **Dice Score Coefficient**: DSC

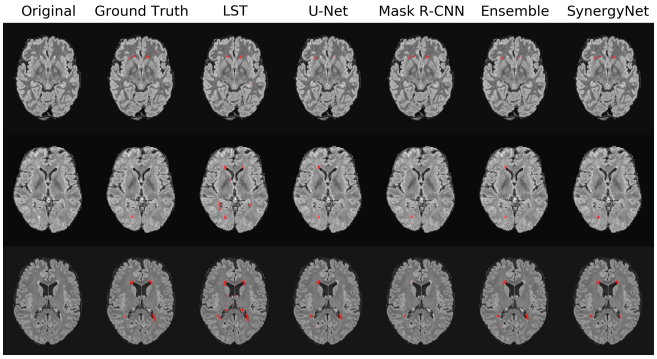

Fig. 3: Comparison of segmentation results from the test set.

Table 1: Performance results for public ISBI-2015 Dataset

Models	DSC	AP	LTPR	LFPR	Sn
U-Net	.5452	.3618	.3435	**.1091**	.4138
Mask R-CNN	.4596	.3145	.4156	.4565	.4368
Ensemble	.5381	.3541	.3632	.1581	.4042
SynergyNet	**.6018**	**.4256**	**.4547**	.2204	**.5067**

Table 2: Performance results for Private MS Dataset

Models	DSC	AP	LTPR	LFPR	Sn
LST (LPA)	.3324	.1758	.3742	.5700	.2762
U-Net	.5999	.4107	.6418	.2334	.5603
Mask R-CNN	.4748	.2802	.5546	.3637	.4043
Ensemble	.5998	.4122	.6378	.2098	.5523
SynergyNet	**.6152**	**.4227**	**.6739**	**.1864**	**.5911**

= 2(O ∩ T)/(| O | + | T |); **Average Precision**: AP summarizes the precision-recall curve and is defined as the weighted mean of precision achieved at each threshold; **Lesion True Positive Rate**: LTPR = $\frac{TPL}{TPL+FNL}$, where TPL and FNL are true positive lesion and false negative lesion respectively; **Lesion False Positive Rate**: LFPR = $\frac{FPL}{TPL+FPL}$, where FPL is false positive lesion; and **Sensitivity**, Sn = $\frac{TP}{TP+FN}$, where TP and FN are voxel true positive and false negative respectively.

We compare our proposed SynergyNet against the lesion prediction algorithm (LPA) [7] as implemented in the LST toolbox version 2.0.15, vanilla U-Net model, vanilla Mask R-CNN model, and the ensemble of these two separately trained models. Images are cropped to a size of 128x160x128 around the image center with random jittering during training. The cropped image is randomly rotated between ± 15 degrees along the sagittal coordinate axis. The Adam optimizer was used with a learning rate of 0.001. The baseline U-Net is trained for 100 epochs and results are reported for the model with the lowest validation loss. We train SynergyNet in the sequential additive manner as previously mentioned.

All the models are tested on the 155 volume held out test set and results are shown in Table 2. Our proposed model score best in all five metrics. The LST algorithm is not competitive against the deep learning methods. With SynergyNet, we see a 2.55%, 2.92%, and 5.00% improvement in Dice score, sensitivity, and LTPR respectively over the baseline vanilla U-Net. Against the ensemble model, the result suggests our fusion approach is advantageous over simply ensembling two separately trained models. This is most likely attributed to the ability of the two specialized branches to inform each other of errors via the shared layers as opposed to no knowledge distillation in the Ensemble approach.

In Figure 2, we see the dice score plotted against the logarithm of the total MRI lesion volumes. SynergyNet's trendline resides above all other trendlines and exhibits the largest gap above the next best trendline on the far left side of the curve. Stratifying the MRIs into lesion volume less than and greater than 2048 voxels, we calculate the mean dice scores for small and large lesion volumes as .5362 and .6954 respectively for U-Net. Similarly for SynergyNet, these dice scores are .5532 and .7052 for small and large lesions respectively. The stratified improvements for these two groups are 3.17% and 1.83% respectively. Coupled with the 5.00% increases in LTPR, this suggests SynergyNet's performance gains are predominantly due to its improved detection and segmentation of small lesions. The performances difference over U-Net tapers off as total lesion volumes approaches 4096 voxels.

Figure 3 shows examples of segmentation outputs from five models. SynergyNet makes fewer and less false positive prediction than U-Net and Ensemble. Where U-Net missed detecting some very small lesions, SynergyNet was able to detect and localize them. LST made many more false positive predictions and sometimes predicts lesion in spurious locations such as inside the cerebrospinal fluid (CSF) area. This is likely due to LST being trained on raw, non-skull stripped FLAIR MRI where the skull and surrounding structure can also be very high in intensity value as well.

For the ISBI public dataset, we were not able to run the LST as the provided FLAIR MRIs were skull-stripped. We ran the remaining four algorithms and show their results in Table 1. SynergyNet scores highest in all but LFPR. We see an improvement of 9.08%-11.81% for the dice score and 35.23%-29.19% for LTPR over vanilla U-Net. U-Net scored highest on LFPR. This is likely due to the data mismatch between the private and public dataset. Lesions in the public dataset are much larger and more confluent, leading to a reduction in the average number of lesions per MRI.

4. CONCLUSIONS

In this paper, we proposed SynergyNet for detection and segmentation of MS lesions in 3D FLAIR MRIs. While U-Net with its global perspective has been favored in many previous studies on MS segmentation, it suffers from poor performance on small lesions. To address this issue, we augment U-Net with a specialized local-focused Mask R-CNN to more reliably detect and segment small lesions. The two network shares many lower layers thus making the entire fusion network memory efficient and able to be trained in an end-to-end fashion. We show our proposed fusion model can improve Dice score and LTPR by 2.55% and 5.0% respectively with reducing LFPR by 20% in a private dataset, and improve Dice score and LTPR by 10% and 32% respectively in the ISBI-2015 public dataset over U-Net. Although our network was designed with MS segmentation in mind, the proposed global-local fusion network is general and can be readily applied to other segmentation tasks.

Acknowledgement The work was partly supported by NVIDIA and NSF grant IIS1715017.

5. REFERENCES

[1] Alastair Compston and Alasdair Coles, "Multiple sclerosis [seminar]," *Lancet*, vol. 9648, pp. 1502–1517, 2008.

[2] Chris H Polman, Stephen C Reingold, Brenda Banwell, Michel Clanet, Jeffrey A Cohen, Massimo Filippi, Kazuo Fujihara, Eva Havrdova, Michael Hutchinson, Ludwig Kappos, et al., "Diagnostic criteria for multiple sclerosis: 2010 revisions to the mcdonald criteria," *Annals of neurology*, vol. 69, no. 2, pp. 292–302, 2011.

[3] Martin Styner, Joohwi Lee, Brian Chin, M Chin, Olivier Commowick, H Tran, S Markovic-Plese, V Jewells, and S Warfield, "3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation," *Midas Journal*, vol. 2008, pp. 1–6, 2008.

[4] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al., "Longitudinal multiple sclerosis lesion segmentation: resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, 2017.

[5] Daniel García-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L Arnold, and D Louis Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," *Medical image analysis*, vol. 17, no. 1, pp. 1–18, 2013.

[6] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson, "Deep learning for brain mri segmentation: state of the art and future directions," *Journal of digital imaging*, vol. 30, no. 4, pp. 449–459, 2017.

[7] Paul Schmidt, *Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging*, Ph.D. thesis, lmu, 2017.

[8] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel, "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation," *Medical Image Analysis*, p. 101557, 2019.

[9] Snehashis Roy, John A Butman, Daniel S Reich, Peter A Calabresi, and Dzung L Pham, "Multiple sclerosis lesion segmentation from brain mri via fully convolutional neural networks," *arXiv preprint arXiv:1803.09172*, 2018.

[10] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159–168, 2017.

[11] Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam, "Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.

[12] Shahab Aslani, Michael Dayan, Loredana Storelli, Massimo Filippi, Vittorio Murino, Maria A Rocca, and Diego Sona, "Multi-branch convolutional neural network for multiple sclerosis lesion segmentation," *NeuroImage*, vol. 196, pp. 1–15, 2019.

[13] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, et al., "Clinically applicable deep learning framework for organs at risk delineation in ct images," *Nature Machine Intelligence*, pp. 1–12, 2019.

[14] Hao Tang, Chupeng Zhang, and Xiaohui Xie, "Nodulenet: Decoupled false positive reductionfor pulmonary nodule detection and segmentation," *arXiv preprint arXiv:1907.11320*, 2019.

[15] Hao Tang, Xingwei Liu, and Xiaohui Xie, "An end-to-end framework for integrated pulmonary nodule detection and false positive reduction," *arXiv preprint arXiv:1903.09880*, 2019.

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

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.