Multitask learning via pseudo-label generation and ensemble prediction for parasitic egg cell detection: IEEE ICIP Challenge 2022

Zaw Htet Aung
Department of Biomedical Engineering, Mahidol University,
Nakhon Pathom,
Thailand
Email: z.zawhtet.a@ieee.org

Kittinan Srithaworn
Looloo Technology,
Bangkok,
Thailand
Email: tunxedge@gmail.com

Titipat Achakulvisut
Department of Biomedical Engineering, Mahidol University,
Nakhon Pathom,
Thailand
Email: titipat.ach@mahidol.edu

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897464

***Abstract*—Parasitic infections are one of the leading causes of deaths and other ailments worldwide. Detecting such infections using traditional diagnostic procedures requires experienced medical technologists together with a significant amount of time and effort. An automated procedure with the ability to accurately detect parasitic diseases can greatly accelerate the process. This work proposes a deep learning-based object detection for parasitic egg detection and classification. We show that multitask learning via pseudo-mask generation improves the single model performance. Moreover, we show that a combination of multitask learning, pseudo-label generation, and ensembling model predictions can accurately detect parasitic egg cells. Continuous training via pseudo-label generation and ensemble predictions improves the accuracy of single-model detection. Our final model achieved a mean precision score (mAP) of 0.956 on a validation set of 1,650 images. Our best model obtained mIoU and mF1 scores of 0.934 and 0.988 respectively. We discuss its technical implementation in this paper.**

***Index Terms*—object detection, pseudo labels, multi-task learning, parasitic egg**

I. Introduction

Despite advances in diagnosis and treatment practices, parasitic infections remain a global health problem to this day. Among them, helminths or intestinal parasites are the most common parasitic infections in the world [1]. Ascaris lumbricoides, one of the many species of helminths, alone is estimated to afflict 1.2 billion humans [2]. Therefore, efficient, robust, and sensitive diagnostic methods are critical for the management of these infections. Conventional light microscopy examination of fecal specimens is one the most commonly used methods for parasitic identification. Due to its ease of use and low cost, it is especially suitable for developing countries where the population's vulnerability to such diseases is significantly higher. However, traditionally, determining the presence or absence of parasitic eggs in a fecal smear is entirely based on the expertise of the medical personnel. A high volume of infections and a subsequent increase in workload deteriorate the efficacy of manual detection. Therefore, an automated pipeline that can accurately localize and classify parasitic eggs from microscopic images is highly desirable.

Numerous classical machine learning and image processing techniques have been proposed to detect parasitic infections in various imaging modalities. Previous methods commonly employ multi-stage detection processes where preprocessing and handcrafted feature extraction are accomplished, followed by classical pattern recognition techniques. In [3], digital image filtering combined with morphometric features is used to train a neural network to identify helminth eggs in microscopic images. Also, a fuzzy expert system based on image moments was proposed by [4] to detect human parasites. Similarly, [5] uses a support vector machine to classify parasites in microscopic images using hand-engineered features. [6] proposed to perform image segmentation of the parasites followed by construction of an object descriptor. Standard pattern classifiers were then used for classification. However, microscopic images of real-world specimens are taken under nonideal conditions, including varying amounts of background debris, imaging artifacts, defocus, and motion blur that limits the performance of handcrafted features.

With the advent of more powerful computing devices and the availability of large datasets, deep learning techniques have begun to thrive recently. They have shown more robustness in deploying under real-world conditions. For example, [7] proposed a convolutional neural network (CNN) with a U-Net structure to segment and classify parasites from microscopic images. A hybrid recurrent and CNN architecture was utilized for classifying parasitic infections [8]. Multiple researches have also treated the diagnosis of parasitic infections as an object detection problem. For example, the FasterRCNN [9] model was trained end-to-end to detect worm eggs in microscopic images. [10] also demonstrated that object detection architecture such as YOLOv5 could be successfully trained to detect parasites in fecal smears. Similarly, [11] proposed FecalNet which used a RetinaNet [12] like architecture to detect fecal components in addition to parasitic eggs. Lastly, [13] proposed the Helminth Egg Analysis Platform (HEAP) which combined a variety of common object detection models in a web-based tool for the diagnosis of helminth infections.

While the usefulness of deep learning-based parasite detection methods has been demonstrated as seen above, the absence of standardized benchmark datasets for this task makes it difficult to compare the efficacy of different algorithms. Therefore, combining benchmark datasets and deep learning techniques can greatly accelerate the research on detection and classification of parasitic cells.

In this paper, we apply object detection techniques to detect and classify parasitic egg cells in the International Conference on Image Processing (ICIP) challenge dataset [14]. We use multiple state-of-the-art CNN models such as high-resolution network (HRNet) [15], ResNet-101 [16], ResNeXt-101 [17] as backbone networks. We apply multi-task learning by training instance segmentation models with pseudo-ground truth masks. These masks are generated once by a class-agnostic instance segmentation model as additional supervisory signals to improve performance. Next, we ensemble predictions from multiple single and multi-task learning architectures to detect parasitic egg cells and to generate pseudolabels for the testing set. Different bounding box predictions are combined with a weighted box fusion algorithm [18]. We then used these pseudo-labels from the testing set and continued training the models for multiple rounds. We found that combining the proposed techniques outperforms a single model prediction. Our final model achieved a mean intersection over union (mIoU) score of 0.934 and a mean F1-Score of 0.988 on the submission leaderboard. We describe its technical implementation in this paper.

II. Materials and methods

We treat the parasitic egg cell detection as an object detection task. Most deep learning-based object detection tasks can be considered as assigning correct category labels to a set of given bounding boxes. The problem can generally be divided into two different approaches: one-stage detection, which learns to detect and classify parasitic egg cells at the same time using prior bounding boxes, and two-stage detection, which first learns to delineate between objects and background using the region of interest (ROI) proposals, and then classify parasitic egg cells. We use anchor-free detectors that predict offsets without anchors, such as Task-aligned One-stage Object Detection (TOOD) [19], and Generalized Focal Loss (GFL) [20] and two-stage detection models including CascadeRCNN [21], and Hybrid Task Cascade (HTC) [22]. We trained and tested the models on the given dataset. Next, we combined their predictions with ensemble techniques. The resulting pseudolabels were then used for continuous training (or fine-tuning). We describe the dataset and our training procedure (Figure 1) in the section below.

A. Datasets

We use parasitic egg detection and classification in the microscopic images dataset from ICIP 2022 challenge [14]. The dataset contains 11,000 images with 11,031 groundtruth bounding boxes. We reserve 15% of images for validation. The square root of annotation areas is distributed around 212.4 ±98.72 pixels. 5.69% of annotations have a square root of an area larger than 400. 742 and 10,286 annotations are classified as medium and large annotations respectively according to the COCO classification standard [23].

B. Training Procedure

1) Pseudo-mask generation for multitask learning via class-agnostic instance segmentation: In our experiment, we employed a class-agnostic instance segmentation model whose mask prediction branch is trained exclusively with ground truth boxes instead of combining them with proposals generated by a region proposal network. This significantly improves the generalization ability of the mask-head which allows the model to predict segmentation masks for novel classes when bounding boxes are provided [24]. We exploit this ability to generate pseudo-ground truth masks for the training and test datasets using the ground truth boxes and the models' predictions respectively. We found that adding pseudo-ground truth masks and retraining the proposed deep learning architectures perform better on the validation and test sets.

2) High-resolution backbone: Being able to provide semantically rich high-resolution representations to detection and instance segmentation heads allows for a more precise spatial localization ability. We found experimentally that maintaining high-resolution feature maps is even more crucial than using multi-scale training or higher resolution input images. Therefore, we opted to use HRNet [15], which generates high-resolution features through repeated multi-resolution fusion as the backbone network. Unlike other models that simply upscale low-resolution representations to create high-resolution counterparts, both are complementary to each other in HRNet.

3) Multi-task learning: Multi-task learning has been shown to improve the performance of deep neural networks especially when training objectives are closely related [25] [26], [27]. For example, object detection and instance segmentation tasks can be learned together in models such as MaskRCNN [26], DETR [28], or Hybrid Task Cascade [22]. Multi-task learning with cascading and interleaving detection and mask prediction improves performances in both tasks for the given dataset. However, annotating ground-truth masks is a much more time-consuming endeavor than the bounding box. Therefore, the availability of datasets with annotated masks is much more limited compared to that of the bounding box dataset.

Multitask cascaded architectures such as CascadeR-CNN [21] and CascadeMaskRCNN [21] are generally more accurate than their single-head counterparts due to the repeated refinement of predictions. Yet, one of the caveats of multi-tasks cascades using CascadeMaskRCNN is the lack of direct communication between mask and bounding box heads. This prevents complementary information flow from each task-specific head and the supervisory signals to be fully utilized by the model. Hybrid Task Cascade (HTC) [22] can mitigate this by adding (i) interleaved bounding box and mask prediction and (ii) a series of connections between mask heads and (iii) spatial context through an auxiliary segmentation branch. Our

proposed method follows the HTC architecture without the semantic segmentation branch as shown in Figure 2.

4) Ensemble predictions: We collect the best performing models and apply ensemble bounding boxes using the weighted box fusion (WBF). WBF combines the confidence scores of proposed bounding boxes using an iterative algorithm to construct average bounding boxes. We use a combination of non-maximum suppression and classification scores thresholding to prune out low-quality predictions. Then, we remove the overlapping boxes by taking the highest confidence box.

5) Pseudo-label generation : Semi-supervised object detection [29] [30], [31] relies on supervisory signals from pseudo labels containing bounding boxes and their corresponding classes. These labels are generated by inferencing unlabeled images with a trained teacher model. We adopt a similar semi-supervised detection pipeline as shown in Figure 1. In addition to pseudo-bounding boxes from the unlabeled testing set, we also generate pseudo masks using the same approach as described above.

C. Training strategy

We train our models with a batch size of 8 and stochastic gradient descent with an initial learning rate of 0.02. The momentum is set to 0.9 and weight decay is set to 10^{-5}. The best performing model in our experiment uses HRNetV2p-W32 as a backbone which has a base channel width (W) of 32 and a feature pyramid (p) style representation for the downstream tasks. The weights are initialized from the model which is pretrained on the COCO dataset. The training is set to 10 epochs with a linear learning rate reduction schedule with a drop rate of 0.1 at the 7th, 8th, and 9th epochs.

We use both geometric and pixel augmentations during training such as affine transformations, random flips, random contrast and brightness, and different types of blur including motion blur. The images are resized so that the shortest side is at most 800 pixels and the longest side at most 1333. Models are trained on 2 to 4 A100 GPUs with 8 images per GPU using mixed-precision mode.

Fine-tuning with pseudo labels on the unlabeled data is accomplished using the same model described above but with the addition of a validation subset. The fine-tuning schedule is set to 5 epochs with learning rate reductions at the 3^{rd} and 4^{th} epochs with a factor of 10.

III. Results

In this section, we describe our results from the competition. First, we show that multitask learning with pseudomask generation outperforms single task prediction. Then, we evaluate whether ensemble predictions with pseudo-label generation improve upon a single model prediction. Continuous training with this procedure gives the highest mIoU and mF1 scores on the leaderboard.

A. Multitask learning outperforms a single-task model

First, we tested if multitask learning with additional pseudo-masks helps improve the task by comparing mAP scores during training (Figure 3). We found that adding an instance segmentation branch as an auxiliary task (CascadeMaskRCNN) increases the model performance with the same backbone (CascadeRCNN) during training. CascadeMaskRCNN outperforms the detection-only CascadeRCNN model by improving mAP from 0.929 to 0.932. The gains in mIoU are more significant than mAP (Table I), which may indicate that adding a mask prediction branch allows precise localization of egg cells. Furthermore, by adding the interleaved connections and direct pathways between mask heads and bounding box heads in the Hybrid Task Cascade (HTC) model, we can further increase the scores on both the validation set (Figure 3) and the leaderboard (Table I). Thus, from our experiment, adding both an auxiliary task and pseudo-labels helped improve both mAP and mIoU scores.

B. Ensemble prediction and pseudo-label generation improve single-model prediction

The trained models' predictions on the test set were fused with WBF with equal weights by default. We set the NMS threshold to 0.55 which favors more confident predictions. Subsequently, we used the best ensemble results as pseudo-labels on the testing set. After obtaining pseudo-labels, we retrain these detection models and perform ensemble prediction again. We do this for 3 iterations for our final model. The final ensemble was a fusion of 5 models: TOOD with Resnet101, GFL with Resnet101, CascadeRCNN with HRNet, HTC with HRNet, and HTC with ResNeXt-101. We achieved mIoU and mF1 scores of 0.9339 and 0.9881 on the submission leaderboard, respectively (Table 1).

IV. Error analysis

We explore the output predictions of the best-performing model on the test scoreboard. The model tends to under or over-predict depending on image conditions. Especially in the presence of background debris with similar textural compositions to the eggs, the model is prone to predict false positive cells as shown in Figure 4B. Using a confidence threshold of 0.5, we can predict around 2,318 cells on test images but it does not cover all images. With a confidence threshold of 0.1, we found that the model outputs 2,936 bounding boxes which cover predictions on all 2,200 test images. It over-predicts to 1.334 cells per image where we found on average around 1.002 cells per image on the training dataset. We think further improvements to the model can reduce the false positives cells and improve the test score. Additionally, we noticed that there is a lower distribution of images with small or medium-size cells in the provided dataset. Therefore, improving detection on medium and small cells can also improve the accuracy on the final test dataset.

V. Conclusion

This paper proposed deep learning-based object detection techniques for detecting parasitic egg cells detection using an ICIP Challenge 2022 dataset. By combining multitask

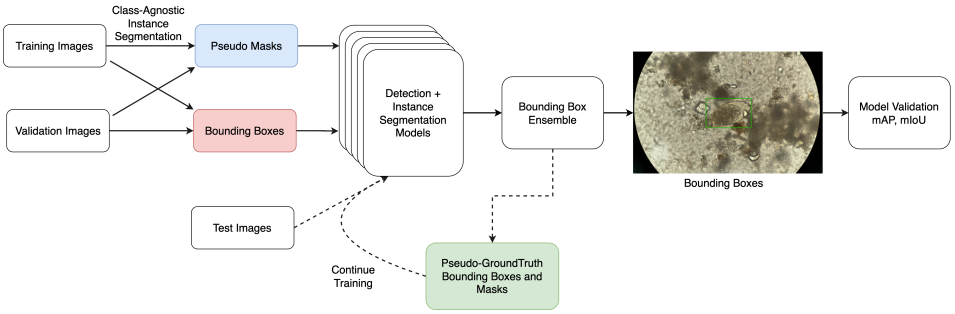


Fig. 1. **Diagram of the training procedure.** First, we used the training data to generate pseudomask labels. Both pseudomasks and bounding boxes are used to train the multitask localization models. We apply the same models to predict pseudolabels and bounding boxes, on the test set and continue training the model. Finally, we validated both mAP and mIoU as our final results.

TABLE I
FINAL MODEL PERFORMANCES AND COMPUTATIONAL COMPLEXITIES

Model	mAP	AP $_{50}$	AP $_{75}$	AP $_{M}$	AP $_{L}$	mIoU	GFLOPs	# Parameters (Millions)
ResNet-101-GFL	0.941	0.998	0.992	0.891	0.941	0.923	215.69	52.33
ResNet-101-TOOD	0.948	0.998	0.993	0.897	0.948	0.926	179.85	53.24
HRNetV2p-W32-CascadeRCNN	0.948	0.999	0.993	0.9	0.948	0.927	312.83	74.71
HRNetV2p-W32-HTC (Multitask)	0.940	1.0	0.998	0.896	0.939	0.928	470.47	82.72
ResNeXt-101-HTC (Multitask)	0.941	1.0	0.999	0.898	0.941	0.928	629.3	139.83
Ensembled Model	**0.956**	0.998	0.993	0.916	0.957	**0.934**	-	-

*Our results show that a single multitask HRNetV2p-W32 model trained with pseudomasks gives the highest mIoU score with less model complexity. Combining ensembling boxes using weighted boxes fusion (WBF) and pseudo-label generation gives the highest performance on the leaderboard.
*AP scores are calculated on the validation set, AP_{50} means average precision when IoU is greater than 50%, AP_{M} means average precision when the size of annotation is medium.
*mIoU scores are from the official scoreboard.
*Multitask means the model is trained with additional pseudomasks.

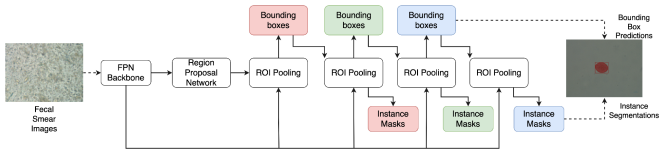


Fig. 2. **Single model architecture.** HRNetV2p-W32 is one of the best-performing single models. The model consists of HRNet backbone with a feature pyramid network (FPN) followed by ROI pooling. We use Hybrid Task Cascade (HTC) as the model head for instance segmentation where it predicts both bounding boxes and masks.

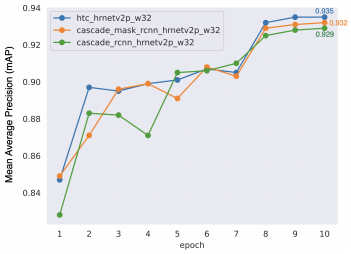


Fig. 3. **Validation mAP (without pseudolabels) scores over training epochs between models .** Multitask learning (CascadeMaskRCNN) outperforms the detection-only (CascadeRCNN) model. Hybrid Task Cascade (HTC) which adds interleaved connections and direct pathways between mask and bounding box heads improves the mAP to 0.935 after 10 epochs.

learning, ensemble prediction, pseudo-label generation, high-resolution representation learning, and cascaded multi-task learning architecture, we show that it can accurately detect parasitic egg cells. Our model gets mIoU and mF1 scores of 0.9339 and 0.9881 respectively.

REFERENCES

[1] "Soil-transmitted helminths," [Accessed 21 April 2022]. [Online]. Available: https://www.euro.who.int/en/health-topics/communicable-diseases/vector-borne-and-parasitic-diseases/soil-transmitted-helminths

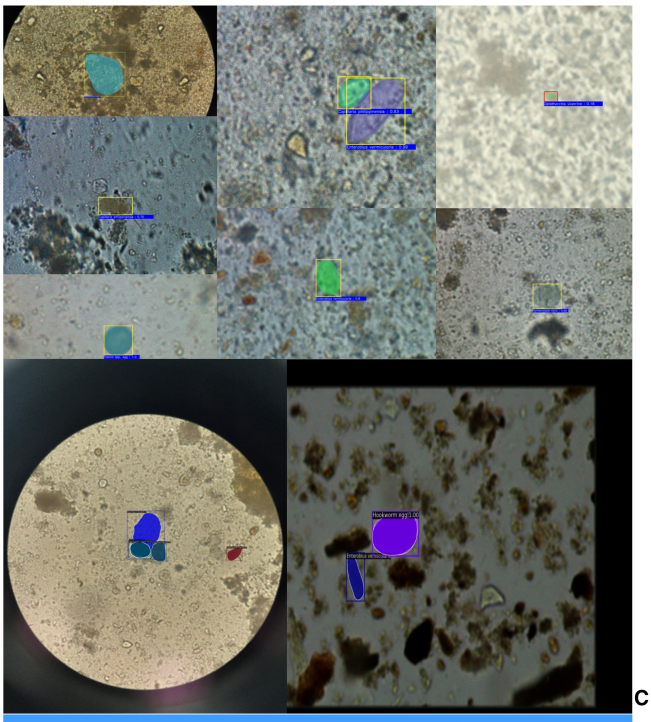
[2] CDC-Centers for Disease Control and Prevention, "CDC - Ascariasis," 2021, [Accessed 21 April 2022]. [Online]. Available: https://www.cdc.gov/parasites/ascariasis/index.html

[3] Y. S. Yang, D. K. Park, H. C. Kim, M.-H Choi, and J.-Y Chai, "Automatic identification of human helminth eggs on microscopic fecal specimens using digital image processing and an artificial neural network," *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 6, pp. 718–730, 2001.

[4] E. Dogantekin, M. Yilmaz, A. Dogantekin, E. Avci, and A. Sengur, "A robust technique based on invariant moments - ANFIS for recognition of human parasite eggs in microscopic images," *Expert Systems with Applications*, vol. 35, no. 3, pp. 728–738, 2008.

[5] D. Avci and A. Varol, "An expert diagnosis system for classification of human parasite eggs based on multi-class SVM," *Expert Systems with Applications*, vol. 36, no. 1, pp. 43–48, January 2009. [Online]. Available: 10.1016/j.eswa.2007.09.012

[6] C. T. N. Suzuki, J. F. Gomes, A. X. Falcao, J. P. Papa, and S. Hoshino-Shimizu, "Automatic Segmentation and Classification of Human Intestinal Parasites From Microscopy Images," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 3, pp. 803–812, March 2013. [Online]. Available: 10.1109/TBME.2012.2187204



Classes	AL	FB	HN	TE	CP	HE	OV	TT	EV	HD	PS
mAP	0.958	0.949	0.940	0.967	0.919	0.923	0.898	0.933	0.934	0.988	0.988

Fig. 4. **Visualization of the predictions of our best model on example test images A**. Example of model predictions on the test set. The model can predict the parasitic egg cells correctly in most cases. B. An example when a model over-predicts on a given image. **B**. The model tends to over predict more parasitic eggs. **C**. False positive examples when a model predicts a dark region as a cell. **D**. Per-class mAP scores of the model on the validation set. *(Please refer to the official dataset for unabbreviated classnames. [14])*

[7] Y. Li, R. Zheng, Y. Wu, K. Chu, Q. Xu, M. Sun, and Z. J. Smith, "A low-cost, automated parasite diagnostic system via a portable, robotic microscope and deep learning," *Journal of Biophotonics*, vol. 12, no. 9, p. e201800410, 2019. [Online]. Available: 10.1002/jbio.201800410

[8] A. Simon, R. Vinayakumar, V. Sowmya, K. P. Soman, and E. A. A. Gopalakrishnan, *Chapter 5 - A deep learning approach for patch-based disease diagnosis from microscopic images," in Classification Techniques for Medical Image Analysis and Computer Aided Diagnosis*, N. Dey, Ed. Academic Press, 2019.

[9] N. Q. Viet, D. T. ThanhTuyen, and T. H. Hoang, "Parasite worm egg automatic detection in microscopy stool image based on Faster R-CNN," in *Proceedings of the 3rd {International} {Conference} on {Machine} {Learning} and {Soft} {Computing}*. New York, NY, USA: Association for Computing Machinery, January 2019, pp. 197–202. [Online]. Available: 10.1145/3310986.3311014

[10] Y. Huo, J. Zhang, X. Du, X. Wang, J. Liu, and L. Liu, "Recognition of parasite eggs in microscopic medical images based on YOLOv5," in *2021 5th {Asian} {Conference} on {Artificial} {Intelligence} {Technology} ({ACAIT})*, October 2021, pp. 123–127. [Online]. Available: 10.1109/ACAIT53529.2021.9731120

[11] Q. Li, S. Li, X. Liu, Z. He, T. Wang, Y. Xu, H. Guan, R. Chen, S. Qi, and F. Wang, "FecalNet: Automated detection of visible components in human feces using deep learning," *Medical Physics*, vol. 47, no. 9, pp. 4212–4222, 2020. [Online]. Available: 10.1002/mp.14352

[12] Tsung-Yi Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv:1708.02002 [cs]*, February 2018.

[13] Chi-Ching Lee, Po-Jung Huang, Yuan-Ming Yeh, Pei-Hsuan Li, Cheng-Hsun Chiu, Wei-Hung Cheng, and P. Tang, "Helminth egg analysis platform (HEAP): An opened platform for microscopic helminth egg identification and quantification based on the integration of deep learning architectures," *Journal of Microbiology, Immunology and Infection*, September 2021. [Online]. Available: 10.1016/j.jmii.2021.07.014

[14] "ICIP 2022 Challenge – Parasitic Egg Detection and Classification in Microscopic Images," [Accessed 21 April 2022]. [Online]. Available: https://icip2022challenge.piclab.ai/

[15] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, October 2021. [Online]. Available: 10.1109/TPAMI.2020.2983686

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 06 2016, pp. 770–778. [Online]. Available: 10.1109/CVPR.2016.90

[17] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 07 2017, pp. 5987–5995. [Online]. Available: 10.1109/CVPR.2017.634

[18] R. A. Solovyev and W. Wang, "Weighted Boxes Fusion: ensembling boxes for object detection models," *CoRR*, vol. abs/1910.13302, 2019. [Online]. Available: http://arxiv.org/abs/1910.13302

[19] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned One-stage Object Detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, oct 2021, pp. 3490–3499. [Online]. Available: 10.1109/ICCV48922.2021.00349

[20] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection," *arXiv:2006.04388 [cs]*, June 2020.

[21] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," *arXiv:1712.00726 [cs]*, December 2017.

[22] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid Task Cascade for Instance Segmentation," *arXiv:1901.07518 [cs]*, April 2019.

[23] Tsung-Yi Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer {Vision} – {ECCV} 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755. [Online]. Available: 10.1007/978-3-319-10602-1_48

[24] V. Birodkar, Z. Lu, S. Li, V. Rathod, and J. Huang, "The surprising impact of mask-head architecture on novel class segmentation," *arXiv:2104.00613 [cs]*, August 2021.

[25] M. Crawshaw, "Multi-Task Learning with Deep Neural Networks: A Survey," *arXiv:2009.09796 [cs, stat]*, September 2020.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv:1703.06870 [cs]*, January 2018.

[27] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking," *International Journal of Computer Vision*, vol. 129, no. 11, pp. 3069–3087, November 2021. [Online]. Available: 10.1007/s11263-021-01513-4

[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *arXiv:2005.12872 [cs]*, May 2020.

[29] K. Sohn, Z. Zhang, Chun-Liang Li, H. Zhang, Chen-Yu Lee, and T. Pfister, "A Simple Semi-Supervised Learning Framework for Object Detection," *arXiv:2005.04757 [cs]*, December 2020.

[30] H. Li, Z. Wu, A. Shrivastava, and L. S. Davis, "Rethinking Pseudo Labels for Semi-Supervised Object Detection," *arXiv:2106.00168 [cs]*, December 2021.

[31] Q. Zhou, C. Yu, Z. Wang, Q. Qian, and H. Li, "Instant-Teaching: An End-to-End Semi-Supervised Object Detection Framework," *arXiv:2103.11402 [cs]*, March 2021.