IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 February 2024, accepted 27 February 2024, date of publication 5 March 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3374073

RESEARCH ARTICLE

Polite Teacher: Semi-Supervised Instance Segmentation With Mutual Learning and Pseudo-Label Thresholding

DOMINIK FILIPIAK[1,2,3,4], ANDRZEJ ZAPAŁA[3], PIOTR TEMPCZYK[1,5,6], ANNA FENSEL[2,7], AND MAREK CYGAN[3]

[1]AI Clearing Inc., Austin, TX 78752, USA
[2]Department of Computer Science, University of Innsbruck, 6020 Innsbruck, Austria
[3]Institute of Informatics, University of Warsaw, 00-927 Warsaw, Poland
[4]Perelyn, 00-388 Warsaw, Poland
[5]NASK National Research Institute, 01-045 Warsaw, Poland
[6]PL4AI-Polish Lab for AI, 00-707 Warsaw, Poland
[7]Artificial Intelligence Chair, Wageningen University and Research, 6708 PB Wageningen, The Netherlands

Corresponding author: Dominik Filipiak (dominik.filipiak@student.uibk.ac.at)

This work was supported in part by Interreg "Österreich-Bayern 2014–2020 Program Project Künstliche Intelligenz (KI)-Net: Bausteine für KI-basierte Optimierungen in der industriellen Fertigung" under Grant AB 292, in part by NVIDIA, in part by Intel, in part by the Polish National Science Center under Grant UMO2017/26/E/ST6/00622, and in part by the European Research Council (ERC) Starting Grant TOTAL. The work of Marek Cygan was supported by the National Centre for Research and Development as a Part of European Union (EU) supported Smart Growth Operational Program 2014–2020 under Grant POIR.01.01.01-00-0392/17-00.

ABSTRACT We present Polite Teacher, a simple yet effective method for the task of semi-supervised instance segmentation. The proposed architecture relies on the Teacher-Student mutual learning framework. To filter out noisy pseudo-labels, we use confidence thresholding for bounding boxes and mask scoring for masks. The approach has been tested with CenterMask, a single-stage anchor-free detector. Tested on the COCO 2017 `val` dataset, our architecture significantly (approx. +8 pp. in mask AP) outperforms the baseline at different supervision regimes. To the best of our knowledge, this is one of the first works tackling the problem of semi-supervised instance segmentation and the first one devoted to an anchor-free detector. The code is available: github.com/AI-Clearing/PoliteTeacher.

INDEX TERMS Semi-supervised instance segmentation, anchor-free detection, instance segmentation, semi-supervised learning.

I. INTRODUCTION

The advent of deep learning transformed computer vision pipelines both in academia and industry. However, progress is often hindered, since deep learning models are expensive to train for several reasons. Leaving the hardware and computational expenses aside, the vast share of costs often comes from providing the right amount of samples to learn from. For a number of supervised problems in computer vision, it is relatively easy to obtain data. However, labelling them is often the real source of expenses. Creating pixel-wise annotations is a tedious and time-consuming process compared to image-level annotations. While this does not mean that the problem scales proportionally with the image size (methods facilitating labelling is a separate subject of research), in practice, it often makes it slower by at least one order of magnitude. The requirement of meticulous data inspections hampers applying machine learning in domains with high-resolution images. Moreover, a number of domains (such as some aerial or medical data) require very specific domain knowledge from labellers, which makes it impossible to easily speed up the process by hiring more labellers. Therefore, an intense effort has been observed in the area of label-efficient machine learning. Semi-supervised learning

The associate editor coordinating the review of this manuscript and approving it for publication was Amin Zehtabian.

methods are tailored to deal with the situation in which there are *enough* data samples, but access to the labels is severely limited.

Semantic segmentation (sometimes called dense classification) is a classical computer vision task of assigning each pixel a category. This enables clustering images into semantically coherent parts. Object detection is concerned with the location and identification of semantic objects on images. Instance segmentation combines these two – it is concerned with locating and identifying entities with pixel-wise accuracy. While an intense research activity can be observed in the areas of semi-supervised semantic segmentation [1], [2] and object detection [3], [4], very little attention has been devoted to semi-supervised instance segmentation methods in the computer vision community.

We propose Polite Teacher, a simple yet effective method for the task of semi-supervised instance segmentation. The architecture is built on the Teacher-Student framework. Its *politeness* in the name is an acronym for *pseudo-label thresholding*, which is concerned with filtering noisy pseudo-labels in detection and mask heads. Our contributions can be summarised as follows:

- We present Polite Teacher – one of the first works devoted to semi-supervised instance segmentation and the first one devoted to modern anchor-free detectors. Our approach uses mask scoring [5] for the pseudo-mask thresholding.
- The presented method significantly (approx. +8 pp. in mask AP) improves baseline performance with different supervision regimes on COCO 2017 and sets the new baseline for further comparison, becoming de facto the new state-of-the-art for this dataset.
- To the best of our knowledge, we are also the first to predict centre-ness in a fashion similar to Mask-scoring R-CNN in order to provide thresholding for bounding boxes.
- As there were no standard benchmarks for semi-supervised semantic segmentation, we also proposed to adopt specific and reproducible regimes and data splits on MS COCO from Unbiased Teacher [3], which was also followed by more recent research [6].

The paper is organised as follows. Section II contains a comprehensive survey of related research. In Section III, we present the details of our method – Polite Teacher. The results of the evaluation are discussed in Section IV, along with the detailed analysis and ablation studies. The paper is concluded with a short summary in Section V.

II. RELATED RESEARCH

This section presents a comprehensive survey of areas adjacent to our work. We present recent research in the area of instance segmentation. Then, we summarise the recent progress in semi-supervised learning. Finally, we combine these two areas and briefly discuss the body of knowledge for semi-supervised instance segmentation.

A. INSTANCE SEGMENTATION

Instance segmentation is a computer vision problem concerned with pixel-wise delineating instances of semantic classes. As it can be perceived as a combination of object detection and semantic segmentation, the advances in instance segmentation are tightly coupled with the two aforementioned tasks (especially the former). In recent years, two kinds of object detectors have been popular: single- and two-staged. A typical example of the latter category is Faster R-CNN [7]. It consists of the backbone feature extractor (eg. ResNet) and two heads: the region proposals network (RPN) and the second one for final detections (RoI – the region of interests head). The proposal candidates are searched on a pre-defined set of anchors using RPN and they are later refined with the RoI head. He et al. [8] introduced Mask R-CNN, which added the mask head to Faster R-CNN to solve segmentation tasks on predicted bounding boxes. Mask scoring [5] adds another head on top of that – it regresses the IoU (intersection over union) score of the predicted masks to improve model robustness. Single-stage detectors try to achieve the outcomes of the aforementioned architecture in a single pass. This often results in higher speed at the expense of precision. Notable examples of such detectors include the YOLO family [9] or RetinaNet [10]. More recently, Lee and Park [11] proposed CenterMask, an anchor-free instance segmentation framework targeted at real-time applications. It is built on FCOS [12], which details are discussed in Section III. The architecture of CenterMask2 introduces spatial attention-guided masks (SAG-Mask) along with backbone feature extractors tailored for instance segmentation. Recently, there has been a surge of research on architectures utilising the concept of self-attention (also called transformers).

Dosovitskiy et al. [13] introduced the visual transformer (ViT), which successfully adapted self-attention to computer vision. This seminal work has sparked research interest in transformers in the vision community. For instance, DINO [14] adapts the Teacher-Student paradigm and self-supervised learning for various vision tasks, such as object detection. MaskDINO [15] added mask prediction to DINO and topped several instance segmentation benchmarks. However, while displaying exceptional performance, solutions purely based on ViT suffer from quadratic computational complexity, which hinders their adoption. Fang et al. proposed EVA [16], a ViT-based foundation model targeted at vision tasks, which was trained on almost 30M images in a self-supervised fashion. While the concept of foundation models is primarily known from the language domain, this approach turned out to yield competitive results on downstream vision tasks, including semantic segmentation. Notably, the finetuned EVA paired with the CMask R-CNN [17] detector outperformed e.g. MaskDINO on MS COCO and set new state-of-the-art LVIS datasets. However, it is Co-DETR [18] which holds the current best result on MS COCO. DETR [19] leveraged ViTs and framed the detection problem

as a direct set prediction. Co-DETR can be viewed as a variant of DETR [19] with an optimised training scheme.

B. SEMI-SUPERVISED LEARNING

Semi-supervised learning techniques can be framed as a middle ground between supervised and unsupervised learning since data with and without labels participate in the learning process. It is related to weakly supervised and self-supervised learning. Some approaches to the problem of semi-supervised learning such as Γ model [20], Π model or temporal ensembling [21] use the notion of self-ensembling. However, more modern ones are focused on the non-standard architecture during the training phase, often incorporating multiple subnetworks. Following Peláez-Vegas et al. [22], one can distinguish several types of modern semi-supervised semantic segmentation approaches: pseudo-labelling, consistency regularisation, adversarial methods, pseudo-labelling, and contrastive learning. While this paper is concerned with instance segmentation, we will adapt this taxonomy for the purpose of this section. With that said, many modern methods often can be classified into several of these categories.

Pseudo-labelling can be perceived as an intuitive and straightforward approach, in which one model is trained on the labelled data, and then that model is used to generate pseudo-labels for another model. Consistency regularisation methods focus on utilising the smoothness assumption and different perturbations to the images, features and networks. For instance, Tarvainen and Valpola [1] introduced Mean Teacher, which is a popular semi-supervised training framework utilising pseudo-labelling and consistency regularisation. It overcomes the limitations of Temporal Ensembling and Π models. Instead of using the standard gradient-based approach, the teacher is updated using the exponential moving average (EMA). Unbiased Teacher [3] builds on top of the Mean Teacher framework – it does add focal loss and confidence thresholding of pseudo labels. Focal loss borrowed from the work of Lin et al. [10] helps with the class imbalance, whereas confidence bounding box thresholding reduces the influence of noisy pseudo-labels. The recent Unbiased Teacher v2 [4] extends it to anchor-free detectors and tackles the issue of the pseudo-labelling on bounding box regression. Besides the Teacher-Student paradigm, there are also other approaches. Cross pseudo supervision for semantic segmentation [2] is another example of a consistency regularisation method. Here, two networks are trained on the output of each other and are penalised for discrepancies in predictions.

As for the other methods, generative ones utilise adversarial training introduced by Goodfellow et al. in their seminal work [23]. While such an approach has been tested and proven to increase the performance in a fully-supervised instance segmentation setting (e.g. [24]), they have not been adapted to a semi-supervised regime yet. Contrastive methods constitute another approach to semi-supervised learning, in which the loss function is shaped to promote placing similar samples close to each other in the resulting feature space. For the dissimilar ones, this is reversed – the loss is penalised if they are close. Regional contrast, abbreviated as ReCo [25] belongs to this category. While using the Teacher-Student framework, this model introduces a dedicated loss function and utilises the semantic relationship between classes.

C. SEMI-SUPERVISED INSTANCE SEGMENTATION

Contrary to object detection and semantic segmentation, instance segmentation in the semi-supervised setting received little attention among scholars so far. Concurrently to our work, Wang et al. [26] presented Noisy Boundaries (NB). This framework also uses the Teacher-Student paradigm and introduces different bounding box thresholds per category, drawing from the work of Radosavovic et al. [27]. The NB architecture has also two special features: the noise-tolerant mask head and boundary-preserving re-weighting. While the noise-tolerant head works with low-level resolution features to suppress the noise on mask boundaries, the boundary-preserving map is focused on highlighting the boundary region for the segmentation part. At the time of writing this publication, the problem of semi-supervised instance segmentation with anchor-free detectors has never been tackled in the literature. After the submission of this paper, Berrada et al. [6] proposed tackling the problem with guided distillation. Their approach resembles Polite Teacherat its core, albeit several improvements have been introduced. Notably, a new guided burn-in phase utilising both labelled and unlabelled samples allowed us to reach new state-of-the-art in semi-supervised instance segmentation. They also evaluated their approach not only within the Mask-RCNN framework but also with more modern Mask2Former [28] on ViT-based backbones.

III. POLITE TEACHER

This section is devoted to the introduction of Polite Teacher. First, we formulate the problem we are solving – semi-supervised instance segmentation. Then, we introduce the architecture of our solution – used detectors, the Teacher-Student learning paradigm, and pseudo-label thresholding. The section concludes with a detailed description of the used loss function.

A. PROBLEM FORMULATION

We consider the problem of *semi-supervised instance segmentation*. Instance segmentation is a computer vision task which combines object detection and semantic segmentation. Semi-supervised setting means that only part of the data available during the training phase is labelled. More formally, we consider training dataset $\mathcal{D}$ consisting of a set of N_{sup} labelled ($\mathcal{D}^{\text{sup}} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_{\text{sup}}}$) and N_{unsup} unlabelled ($\mathcal{D}^{\text{unsup}} = \{\mathbf{x}_i\}_{i=1}^{N_{\text{unsup}}}$) images. Here, $\mathbf{x}_i$ and $\mathbf{y}_i$ stand for images and their labels (instances categories along with

their bounding boxes and masks) respectively. Typically, $N_{\texttt{unsup}} \gg N_{\texttt{sup}}$. In this work, we assume that $\mathcal{D}^{\text{sup}}$ and $\mathcal{D}^{\text{unsup}}$ come from the same distribution.

B. ARCHITECTURE

The architecture of Polite Teacher depends on several components. The first one is the detector, which is used twice due to the Teacher-Student paradigm. We use CenterMask [11]. This is a single-stage anchor-free detector, which has a relatively simple architecture and therefore is easy to tune. Two such networks are then framed in the Teacher-Student paradigm to handle both labelled and unlabelled data. Finally, two-fold pseudo-label thresholding takes place to remove noisy ones. The first one uses bounding box uncertainty, and the second one rejects masks with an estimated low IoU score.

C. DETECTOR

To properly present CenterMask, FCOS should be discussed first. Tian et al. [12] introduced *Fully Convolutional One-Stage Object Detector* (abbreviated as FCOS), an anchor-free object detector. In general, one-stage detectors due to the lack of the proposal generation phase have fewer hyper-parameters to tune and therefore they are easier to train. Being anchor-free means eliminating pre-defined anchors, which diminishes the computational burden related to calculating IoU scores. FCOS frames detection as a per-pixel prediction task, which resembles semantic segmentation. Three loss components are subject to optimisation: classification, regression, and centre-ness. While classification works similarly to other detectors, the regression targets are quite different. Instead of predicting bounding box corners (like in e.g. Faster R-CNN), the four regressed values are l (the distance from the centre of a bounding box to its left border), t (top), r (right), b (bottom). Finally, the centre-ness denotes the centre of a given bounding box. Ground-truth centre-ness for (l^*, t^*, r^*, b^*) is defined as follows:

$$\text{centreness}^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}}. \tag{1}$$

Intuitively, this approach promotes bounding boxes which are located at the centre of a given object.

Lee and Park [11] introduced CenterMask, which extends FCOS for the task of instance segmentation. This is done similarly to how Faster R-CNN [7] is extended by Mask R-CNN [8]. However, there are some differences. For instance, the RoI assignment function is redefined due to the different levels of the feature pyramid (FPN) which are used. Instead of the mask head from Mask R-CNN, CenterMask utilises the spatial attention-guided mask (abbreviated as SAG-Mask). For $\mathbf{x}$, which here mean features extracted from RoI align, the attention-guided feature map is calculated as follows:

$$\mathbf{x}_{\text{sag}} = \sigma\left(\text{conv}_{3\times3}\left(\text{concat}\left(P_{\max}, P_{\text{avg}}\right)\right)\right) \odot \mathbf{x}, \tag{2}$$

where σ denotes sigmoid function, $\text{conv}_{3\times3}$ is convolutional layer with 3×3 filter, $P_{\max}$ and P_{avg} are the results of max and average pooling, and concat stands for the concatenation.

D. TEACHER-STUDENT LEARNING

We adopt a 2-step training procedure. In the first step, the model is trained using only labelled data ($\mathcal{D}^{\text{sup}}$), which makes this part a standard supervised instance segmentation. Instead of using a fixed number of batches for this step – as *burn-in stage* in Unbiased Teacher [3] – we rather train it as long as it converges in terms of mask AP and take the best model θ to ensure the highest results. Naturally, this step is expected to take longer with a higher number of supervised examples. In the second step, mutual Teacher-Student learning with pseudo-labels takes part. The best model from the first step is used and copied to be used as student and teacher models ($\theta_s \leftarrow \theta, \theta_t \leftarrow \theta$). The model can be trained with the burn-in stage as well.

Teacher and student models receive the same input data – they are augmented differently, though. The teacher receives moderately augmented images (*weak* augmentations – we use random flipping), whereas the student consumes visibly perturbed images (*strong* augmentations – same as weak, plus colour jitter, random grayscale, gaussian blur, and random patch erasing). During the training, the predictions from the teacher model serve as pseudo-labels (bounding boxes with their classes and masks) for the student network. The teacher is updated using the exponential moving average – see equations 6 and 7 in the next subsection. Figure 1 illustrates the process.

E. PSEUDO-LABEL THRESHOLDING

As the teacher is used to generate pseudo-labels $\hat{\mathbf{y}}$ in the semi-supervision regime, they can be noisy – especially with a high share of unsupervised data. Therefore, Polite Teacher uses two-step pseudo-label thresholding: one is concerned with bounding boxes, whereas the second one refines the masks. Similarly to STAC [29] and Unbiased Teacher [3], we introduce a bounding box confidence threshold – τ_{cls}. Bounding boxes with a classification score smaller than τ_{cls} are discarded and not used further in the training. The sigmoid output of the classification is treated here as confidence. Inspired by the work of Huang et al. [5], we also use a mask-scoring mechanism. It regresses the IoU values of the generated masks and improves instance segmentation performance due to the prioritisation of more accurate masks. While not directly designed for the task of semi-supervised learning, the output of this block can be used for straightforwardly filtering noisy pseudo-masks. That is, only masks satisfying $\hat{\mathbf{y}}_{\text{IoU}} > \tau_{\text{IoU}}$ are used in the unsupervised learning stage. The other ones are considered uncertain and receive zero gradients.

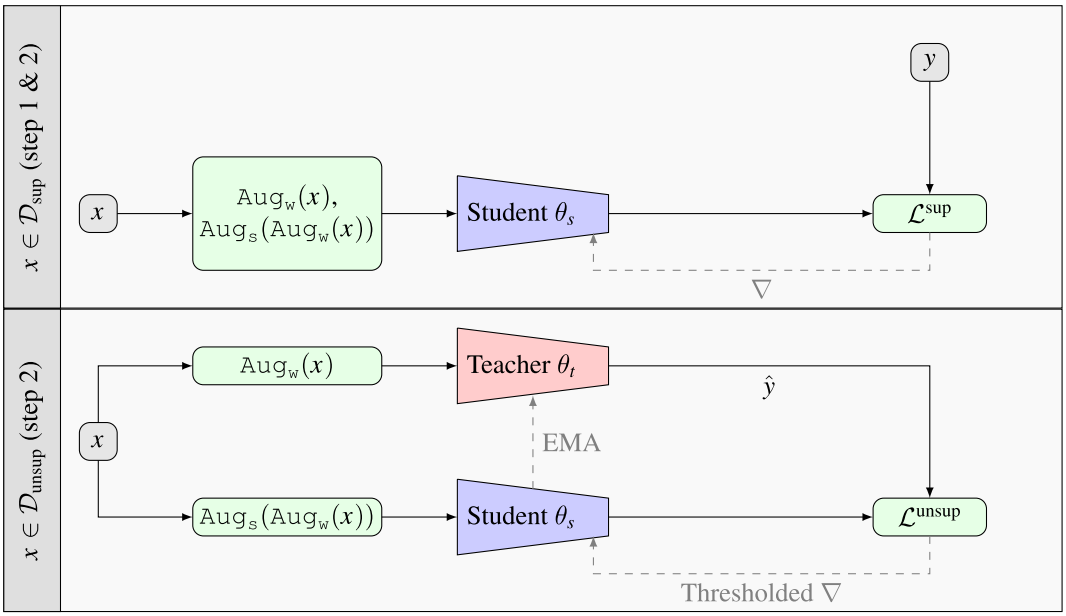

FIGURE 1. Architecture and data flow of Polite Teacher with regard to the supervised and unsupervised data handling. Aug_s and Aug_w represent strong and weak augmentations respectively.

F. OPTIMISATION

The overall batch-wise loss $\mathcal{L}$ for supervised $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{B_{\text{sup}}}$ and unsupervised $\left\{\left(\mathbf{x}_j, \hat{\mathbf{y}}_j\right)\right\}_{j=1}^{B_{\text{unsup}}}$ examples in a batch is computed as follows:

$$\mathcal{L} = \sum_{i}^{B_{\text{sup}}} \mathcal{L}^{\text{sup}}\left(\mathbf{x}_i, \mathbf{y}_i\right) + \lambda \sum_{j}^{B_{\text{unsup}}} \mathcal{L}^{\text{unsup}}\left(\mathbf{x}_j, \hat{\mathbf{y}}_j\right), \tag{3}$$

where $\mathcal{L}^{\text{sup}}$ is the loss of the supervised part, and $\mathcal{L}^{\text{unsup}}$ is the loss of the unsupervised part. The unsupervised part is scaled by λ. The supervised component is calculated as follows:

$$\begin{aligned}\mathcal{L}^{\text{sup}}(\mathbf{x}, \mathbf{y}) = \mathcal{L}^{\text{sup}}_{\text{cls}}(\mathbf{x}, \mathbf{y}) + \mathcal{L}^{\text{sup}}_{\text{centre}}(\mathbf{x}, \mathbf{y}) \\ + \mathcal{L}^{\text{sup}}_{\text{box}}(\mathbf{x}, \mathbf{y}) + \mathcal{L}^{\text{sup}}_{\text{mask}}(\mathbf{x}, \mathbf{y}) \\ + \mathcal{L}^{\text{sup}}_{\text{IoU}}(\mathbf{x}, \mathbf{y}),\end{aligned} \tag{4}$$

where $\mathcal{L}^{\text{sup}}_{\text{cls}}$ is the supervised classification loss, $\mathcal{L}^{\text{sup}}_{\text{centre}}$ is the supervised centreness loss, $\mathcal{L}^{\text{sup}}_{\text{box}}$ represents the supervised bounding box regression loss, and $\mathcal{L}^{\text{sup}}_{\text{mask}}$ is the supervised segmentation mask loss, and $\mathcal{L}^{\text{sup}}_{\text{mask\_IoU}}$ is the supervised segmentation mask scoring loss. Regarding the pseudo-labelling loss, we use the following definition:

$$\begin{aligned}\mathcal{L}^{\text{unsup}}\left(\mathbf{x}, \hat{\mathbf{y}}\right) = \mathbb{1}_{\hat{\mathbf{y}}_{\text{cls}} > \tau_{\text{cls}}} \mathcal{L}^{\text{unsup}}_{\text{cls}}\left(\mathbf{x}, \hat{\mathbf{y}}\right) \\ + \mathbb{1}_{\hat{\mathbf{y}}_{\text{IoU}} > \tau_{\text{IoU}}} \mathcal{L}^{\text{unsup}}_{\text{mask}}\left(\mathbf{x}, \hat{\mathbf{y}}\right) \\ + \mathcal{L}^{\text{unsup}}_{\text{IoU}}\left(\mathbf{x}, \hat{\mathbf{y}}\right),\end{aligned} \tag{5}$$

where $\mathcal{L}^{\text{unsup}}_{\text{cls}}$ is the unsupervised classification loss, $\mathcal{L}^{\text{unsup}}_{\text{mask}}$ is the unsupervised segmentation mask loss, and $\mathcal{L}^{\text{unsup}}_{\text{IoU}}$ is the unsupervised segmentation mask scoring loss. Regarding the particular loss functions implementation, we follow FCOS and CenterMask: $\mathcal{L}^{\text{sup}}_{\text{cls}}$ is focal loss [10], $\mathcal{L}^{\text{sup}}_{\text{box}}$ is UnitBox IoU loss [30], $\mathcal{L}^{\text{sup}}_{\text{centre}}$ is binary cross-entropy loss, $\mathcal{L}^{\text{sup}}_{\text{mask}}$ is average binary cross-entropy loss [8], and $\mathcal{L}^{\text{sup}}_{\text{IoU}}$ is L_2 loss. The same losses are used for unsupervised components (where applicable).

The student is trained using a standard stochastic gradient descent, whereas the teacher can be perceived as an ensemble of the students:

$$\theta_s \leftarrow \theta_s - \gamma \frac{\partial\left(\mathcal{L}^{\text{sup}} + \lambda \mathcal{L}^{\text{unsup}}\right)}{\partial \theta_s}, \tag{6}$$

$$\theta_t \leftarrow \alpha \theta_t + (1 - \alpha)\theta_s. \tag{7}$$

Here θ_t and θ_s represent the teacher and student model parameters respectively and α is the EMA coefficient (a hyperparameter). Following Liu et al. [3], the teacher network trained in such a way is more robust to the sudden changes of decision boundaries caused by the minority classes in batches – especially in the presence of pseudo-labels. An important practical implication of this is the fact that there is no need to store gradients for the teacher model, which reduces GPU memory usage (compared to simply training two models).

G. TIME AND MEMORY COMPLEXITY

Complexity-wise, the presented method displays performance similar to other Teacher-Student methods. Batch-wise complexity behaves differently in the fast-forward and backpropagation phases. Additionally, we need to distinguish the burn-in and the main phase. For the sake of simplicity, we omit e.g. augmentations and pseudo-label filtering. An image in a single batch can be categorised as unlabelled (u) and labelled (l) data, and – subsequently – as weakly (w) and strongly (s) augmented. In other words, a batch should contain $n = n_{u,w} + n_{u,s} + n_{l,w} + n_{l,w}$ images. For the

feed-forward burn-in phase, the complexity is simply $\mathcal{O}(nf)$, where f is the cost of running the batch through the network and $n_{u,w} = n_{u,s} = 0$. After the burn-in, the time complexity can be modelled as $\mathcal{O}\left(n_{u,w}f + (n_{l,w} + n_{l,w})f + n_{u,s}f\right)$. Since student and teacher share their architecture, their cost is equal. All that can be simplified to $\mathcal{O}(nf)$.

The cost of calculating the gradient is sublinear in n. Notice that in the backpropagation phase, the gradients aren't calculated for $n_{u,w}$ samples, as the teacher is updated periodically with EMA. Additionally, not all images are always used, since some of them do not have any label which satisfies the conditions ($\hat{\mathbf{y}}_{\text{cls}} \leq \tau_{\text{cls}}$ and $\hat{\mathbf{y}}_{\text{IoU}} \leq \tau_{\text{IoU}}$). In terms of memory complexity, the dominating factor stems from following the Student-Teacher framework, which boils down to storing two networks in the memory. Since they are not needed at the same time, one can consider keeping only one in the memory at the same time, but this will come at the expense of swapping the model every iteration.

IV. EVALUATION

This section describes the evaluation of Polite Teacher. We start with discussing the training setup, implementation details, and dataset. Then, we present the result of our main experiment, which is followed by the detailed analysis and ablation studies of particular components of Polite Teacher.

A. SETUP

All the experiments were performed either on `a2-highgpu-4g` instances on Google Cloud Platform (4xA100, 40 GB RAM each) or various machines with 8 GPUs (up to 16 GB RAM each) on the proprietary cluster (each of which contained Titan V, RTX 2080 Ti, or Titan X GPUs). Polite Teacher was developed on CenterMask2 and Unbiased Teacher codebase – both built on the Detectron2 framework [31].

We evaluate Polite Teacher on the MS-COCO 2017 dataset [32] using different supervision regimes (1%, 2%, 5% and 10% supervised). The supervised-unsupervised split is taken from the Unbiased Teacher [3] – while it was originally meant to be used for evaluation of semi-supervised object detection, it can be used with our method as well. We report evaluation results on `val` subset, as the test one is not publicly available. The reported metric used in this study is mask mAP (mean average precision, simply called AP later on), which is calculated as an average of AP with IoU thresholds set from 0.5 to 0.95 (with 0.05 intervals). Bounding box AP is also reported for selected experiments.

As base hyperparameters, we used the ones set in CenterMask2. The EMA coefficient α for Teacher learning is set to 0.9996. The models were trained for up to 270,000 batches with stochastic gradient descent. We used batch size 32 (16 supervised and 16 unsupervised samples) with a learning rate $\gamma = 0.006$, weight decay of 0.0001 and momentum of 0.9. Similarly to the CenterMask2, the learning rate has been decreased by a factor of 10 on steps 210,000 and 250,000. However, such a long training was often not necessary, as models overfitted on much earlier stages. These experiments have been early stopped. Regarding the pseudo-label thresholding, we used $\tau_{\text{cls}} = 0.6$ and $\tau_{\text{IoU}} = 0.9$. The unsupervised weight has been set to $\lambda = 2$. More details on the last three values are in Section IV-C. We use ResNet-50 backbone [33] for all the experiments.

B. RESULTS

Table 1 shows results for the main experiment conducted on the MS-COCO 2017 validation dataset. Polite Teacher reached 18.33/22.28/26.46/30.08 mask AP on 1%/2%/5%/10% respectively, which stands for +8.26/+8.82/+8.42/+8.00 pp. change in this metric over the baseline CenterMask2 respectively. Figure 2 presents qualitative results from different models created in this experiment at different levels of supervision.

For the vast majority of our experiments, we thought that our method would be the first one devoted to semi-supervised instance segmentation. The recent Noisy Boundaries (NB) approach [26], a concurrent work to Polite Teacher, is also concerned with this problem and has been evaluated on a similar percentage of supervision on COCO 2017. However, these are different splits. We did not perform direct comparisons, as we were not aware of this work for the majority of our research – we report these results for scientific integrity, though. In general, Noisy Boundaries reported a smaller increase in the mask AP (especially with low supervision), although for fair comparison such claims should be made after running the models on *exactly* the same supervised/unsupervised data splits. It is also unclear how much of this difference can be attributed to the different detectors (a two-staged Mask R-CNN has been used). Following Wang et al. [26], we also report the results for Data Distillation (DD) method [27], which was evaluated jointly with NB. It was developed for the task of *omni-supervised* (known also as *webly-supervised*) learning, a special case of semi-supervised learning in which unlabelled data from the Internet are considered during the training. At the heart of this approach lies the pipeline of different data transformations. The results are later ensembled to provide pseudo-labels.

C. DETAILED ANALYSIS AND ABLATION STUDIES

In this section, a detailed analysis of the influence of hyperparameters and ablation studies is presented. Unless otherwise specified, all the configuration is the same as in Section IV-A. All experiments have been performed with the 5% supervision regime.

1) INFLUENCE OF BOUNDING BOX FILTERING THRESHOLD

In this experiment, we investigate the importance of bounding box filtering thresholds. To separate the influence of sole bounding box filtering, we did not include mask IoU in the optimisation – it is the subject of another experiment. That is, $\mathcal{L}^{\text{sup}}_{\text{IoU}}(\mathbf{x}_i, \mathbf{y}_i)$ and $\mathcal{L}^{\text{unsup}}_{\text{IoU}}(\mathbf{x}_j, \hat{\mathbf{y}}_j)$ has been not taken into account in equations 4 and 5 respectively. As it turns out, even this significantly improves mask AP over baselines. Table 2

TABLE 1. Results with ResNet50 backbone. Oracle results reported by Lee et al. [11]. The results for two-stage detectors are taken from the work of Wang et al. [26]. Notice that it uses a random split of the dataset – in particular, this is different from the one used by us. Therefore, these results cannot be directly compared, but we report them in order to trace the comparison with their baselines.

Architecture	% supervised			
	1	2	5	10
Mask AP, single-stage detectors (oracle: 34.70%), COCO 2017 `val`*, split from [3]*				
CenterMask2 [11]	10.07	13.46	18.04	22.08
Polite Teacher (ours)	18.33 $^{(+8.26)}$	22.28 $^{(+8.82)}$	26.46 $^{(+8.42)}$	30.08 $^{(+8.00)}$
Mask AP, two-stage detectors (oracle: 34.50%), COCO 2017 `val`*, split from [26]*				
Mask R-CNN [8]	3.5	9.4	17.3	22.0
DD [27]	3.8 $^{(+0.30)}$	11.8 $^{(+2.40)}$	20.4 $^{(+3.10)}$	24.2 $^{(+2.20)}$
NB [26]	7.7 $^{(+4.20)}$	16.3 $^{(+6.90)}$	24.9 $^{(+7.60)}$	29.2 $^{(+7.20)}$

(a) 1% supervised (b) 2% supervised (c) 5% supervised (d) 10% supervised

FIGURE 2. Qualitative Polite Teacher results on COCO 2017 `val` with different supervision regimes.

and Figure 3 (left) present AP values for this experiment. The bounding box threshold value with the highest bounding box and mask AP was 0.6. Interestingly, this is a slightly smaller threshold than in the original Unbiased Teacher paper (0.7). The difference might stem from the different neural network architectures (Faster R-CNN vs CenterMask). Note that this experiment used suboptimal $\lambda = 0.75$ from Equation 3 and hence the results are slightly worse compared to the following experiments. This is because the experiment to determine the correct unsupervised loss weight was yet to be carried out at this point.

2) INFLUENCE OF UNSUPERVISED LOSS WEIGHT

We also examine the influence of the weight of unsupervised loss, which is denoted as λ in Equation (3). Figure 3 (centre) and Table 3 present detailed results of this study. Similarly

TABLE 2. Influence of bounding box filtering threshold (5% supervision).

Metric	τ_{cls}						
	0.5	0.55	0.6	0.65	0.7	0.8	0.9
AP (box)	25.77	26.91	27.46	25.67	23.71	21.16	18.73
AP (mask)	24.14	24.97	25.32	23.92	22.25	20.03	17.65

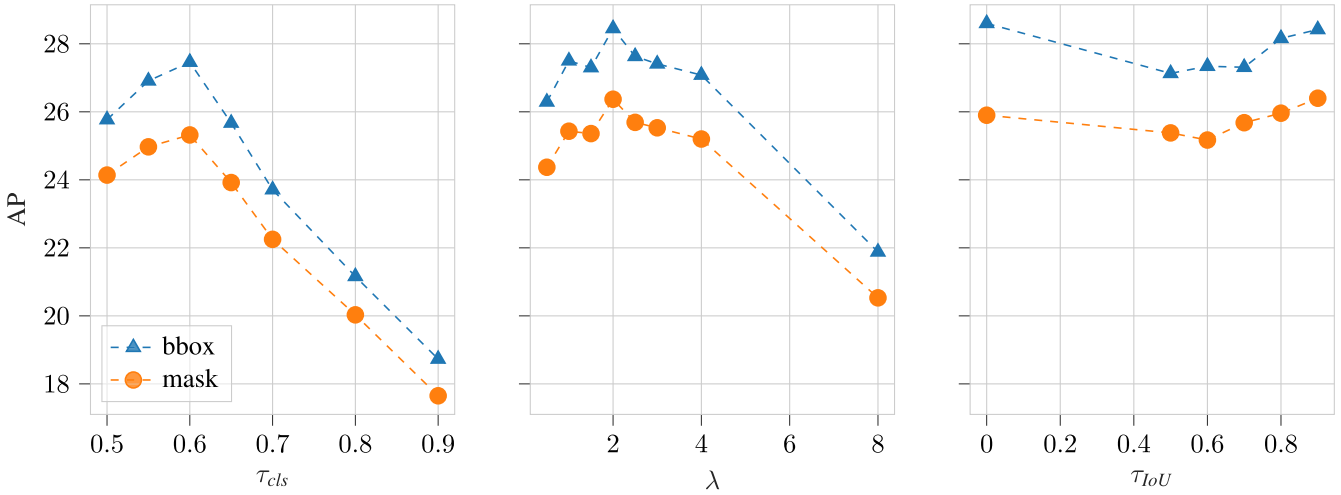

FIGURE 3. Results for experiments controlled for different hyperparameters values in the 5% supervision regime (see Section IV-C, as well as Table 2, 3, and 4).

to the previous experiment, we did not include $\mathcal{L}^{\text{sup}}_{\text{IoU}}(\mathbf{x}_i, \mathbf{y}_i)$ and $\mathcal{L}^{\text{unsup}}_{\text{IoU}}(\mathbf{x}_j, \hat{\mathbf{y}}_j)$ from equations 4 and 5 as the optimisation components. We used $\tau_{\text{bbox}} = 0.6$, which is the result of the previous experiment. The highest mask AP has been obtained at $\lambda = 2.0$. Interestingly, in Unbiased Teacher, which is a similar architecture, this parameter was set to $\lambda = 4.0$.

3) INFLUENCE OF MASK SCORING FILTERING THRESHOLD

In this experiment, we investigate the importance of mask filtering threshold τ_{IoU}. We use $\tau_{\text{cls}} = 0.6$ and $\lambda = 2.0$, as these two values provided the best results in the previous experiments. Figure 3 (right) and Table 4 present the detailed results of this experiment. Compared to the results without mask scoring, the best value ($\tau_{\text{IoU}} = 0.9$) yielded insignificant differences in mask AP (+0.03 pp.) and bounding box AP (−0.03 pp.). Interestingly, the conducted experiment displayed a convex-like U-shaped relationship between τ_{cls} and mask AP. Passing all pseudo-masks resulted in the highest bounding box AP, whereas filtering most of them yielded the highest mask AP.

4) ABLATION ON PSEUDO-BOUNDING BOX THRESHOLDING

For an ablation study, we compare the baseline CenterMask model to the Teacher-Student with bounding box thresholding. Essentially, such a model is very similar to Unbiased Teacher [3], which is proven to greatly improve results for semi-supervised object detection. While raw CenterMask achieved 18.04% on 5% supervision, Polite Teacher yielded 26.46% mask AP, which is a +8.42 pp. increase (see tables 1 and 3). This suggests that much of the mask AP gain can be attributed to the Teacher-Student paradigm with bounding box thresholding.

5) ABLATION ON PSEUDO-MASK THRESHOLDING

In this ablation, we compare the model with bounding box thresholding to the model with bounding box and mask thresholding (that is, Polite Teacher). Judging only by mask AP, the influence of the pseudo-mask filtering threshold on the final results can be easily neglected, as shown in Figure 3 (right). However, applying mask scoring resulted in visibly faster convergence. The model with mask scoring reached 26% mask AP in 40k steps, whereas the model without it needed 74K steps to reach the same value, which is almost two times longer. The highest mask AP values have been reached at 47k (26.39%) and 99k step (26.37%) respectively, which is also close to two times longer. The detailed figures for this run are in tables 3 and 4. In order to check the stability of this behaviour, we repeat these experiments (Figure 4).

6) VARIANCE EXAMINATION

Due to the computational limitations, we are not reporting results as a series of experiments with their means and standard deviations. However, to assess the variance of the proposed method we carried out a separate experiment, in which we ran Polite Teacher training with 5% supervised data several times – each time with a different seed value. Figure 4 presents mean evaluation results per each step

TABLE 3. Importance of unsupervised loss weight (5% supervision).

Metric	λ							
	0.5	1.0	1.5	2.0	2.5	3.0	4.0	8.0
AP (box)	26.29	27.50	27.30	28.45	27.63	27.41	27.08	21.88
AP (mask)	24.37	25.43	25.36	26.37	25.69	25.53	25.20	20.53

TABLE 4. Influence of mask scoring filtering threshold.

Metric	τ_{IoU}					
	0.0	0.5	0.6	0.7	0.8	0.9
AP (box)	28.60	27.13	27.34	27.31	28.16	28.42
AP (mask)	25.90	25.38	25.17	25.68	25.96	26.40

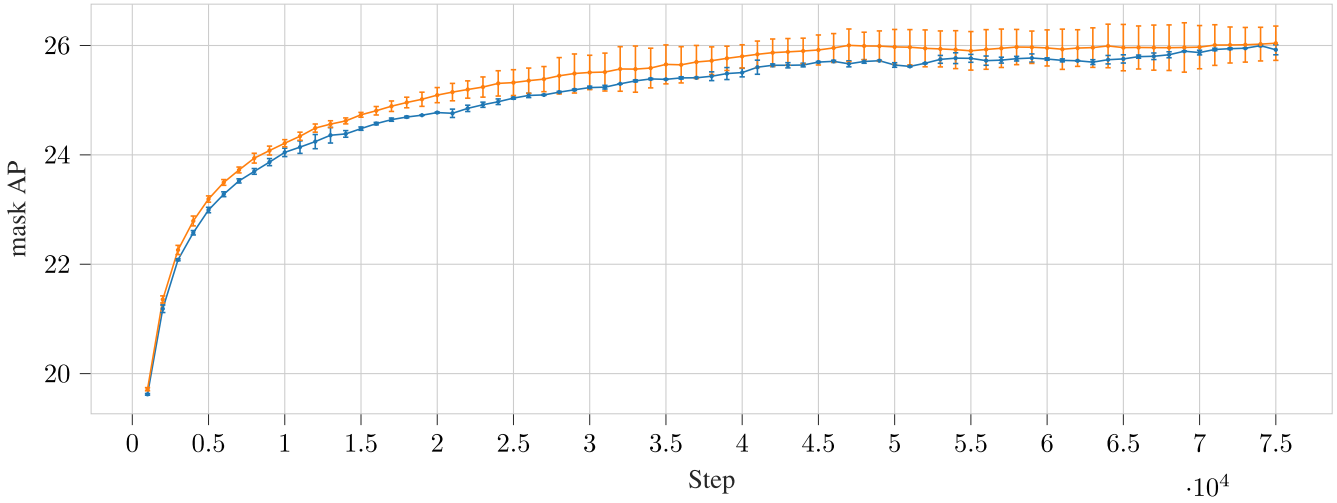

FIGURE 4. Mean results for Polite Teacher training with 5% supervised data, with (orange, 4 runs) and without (blue, 3 runs) mask scoring.

(batch), along with the standard deviations. While the experiment has shown non-homogeneous variance for the variant with mask scoring, the maximum mask AP values are similar: 26.39, 26.39, 26.01, 25.85 with a mean of 26.16 ($\sigma = 0.27$). The variant without mask scoring achieved 26.03, 25.99, 25.99 max mask API (mean 26.00, $\sigma = 0.02$) – that is, an order of magnitude smaller variance, but at the expense of slower convergence and lower metrics value. Interestingly, for the model with the mask scoring head, without the last run, the mean would be 26.26 ($\sigma = 0.22$). While it seems that the last run missed the local optima and much of the per-step variance can be attributed to it, we report all the obtained results for scientific integrity. The high variance might suggest that another hyperparameter should be introduced (e.g. mask scoring head weight) or learning rate should be readjusted.

7) QUALITATIVE ANALYSIS WITH UNEVEN LIGHT CONDITIONS

Dealing with *noisy* input is one of the practical challenges associated with visual models. While the source of noise can be e.g. poor quality of the camera, natural effects, such as uneven light conditions can deteriorate the performance of a good model. Therefore, we performed another series of qualitative tests on pre-selected images with non-uniform illuminance. First, we tested some COCO images in the presence of shadow (Figure 5). In the image with COCO ID `802`, almost half of the oven is covered in shadow. However, all the models handled it very well. Despite different levels of luminosity, images `1268` and `4135` posed no challenge to all the models. This might be because the `people` class is well represented in MS COCO. In `5992`, most models had the problem of separating instances of the same `sheep` class. We also tested some COCO images in the presence of light reflections, which turned out to be an interesting and challenging case. In `2592`, one can see that there is a visible reflection in the blade of the knife. Models failed to notice the knife, the analysis has shown that some knife regions were detected, but the reflection in the middle "broke" the continuity of the prediction and it was eventually suppressed. Similarly, the light reflections on the image with ID `107087` confused most of the models. We also provide more qualitative results with uneven light

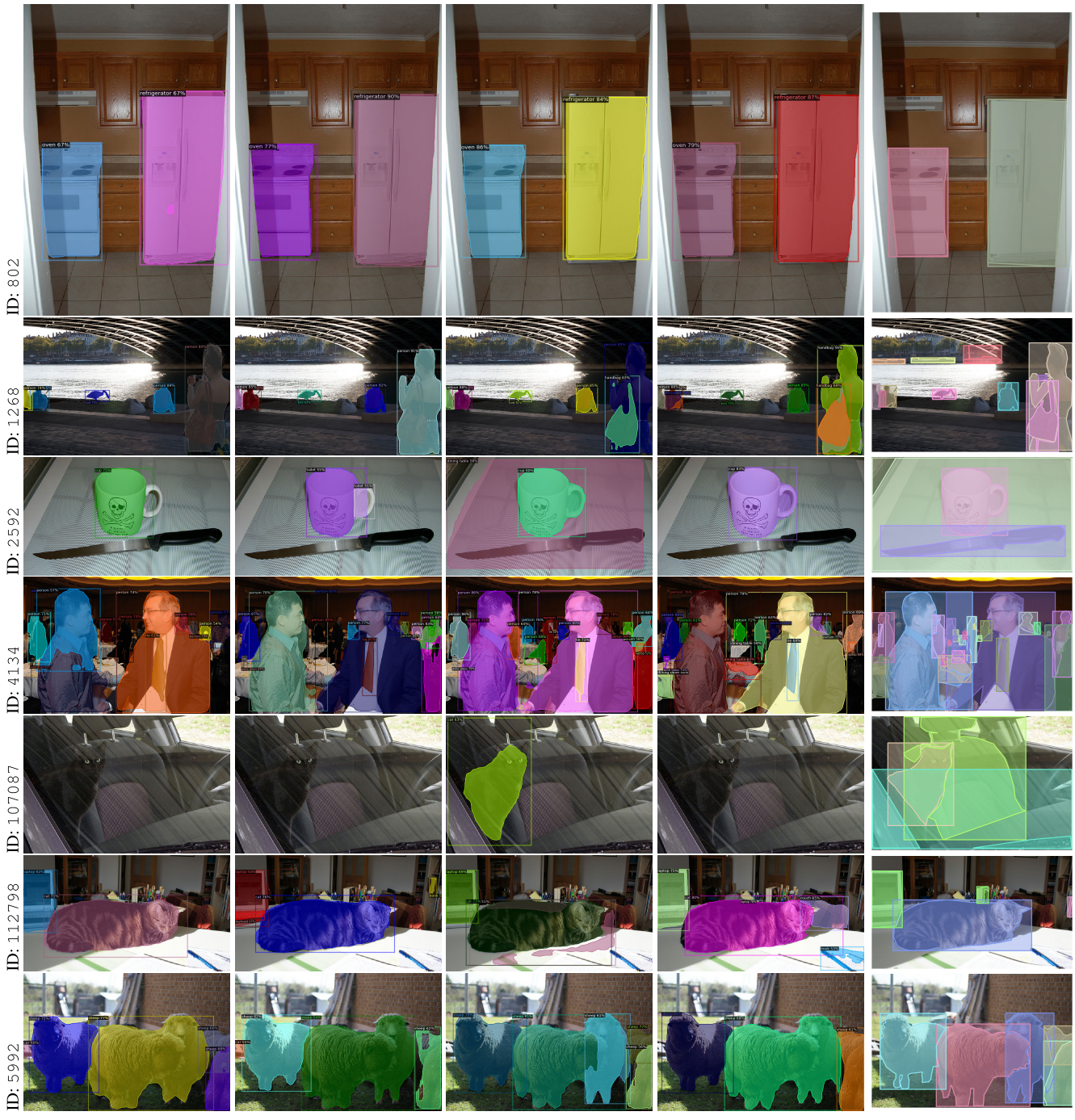

FIGURE 5. Qualitative Polite Teacher results with different supervision regimes on uneven light conditions. Images sources: COCO 2017 `val`.

conditions on the custom images (that is, not from the COCO dataset). Results from Figure 6 show that – contrary to COCO examples – shadow can become a challenge (`CAT-1`), where the beamer light glare did not influence the mask this time (`CAT-2`). Therefore, it is hard to generalise these results to a concise conclusion and more tests are needed in future work.

V. SUMMARY

We presented Polite Teacher, a simple and effective architecture for semi-supervised instance segmentation. Tested with a CenterMask, a single-stage detector, our approach yielded approx. +8 pp. mask AP on different supervision regimes with COCO 2017, while it introduces only three hyperparameters to tune. A certain limitation of this study is the lack

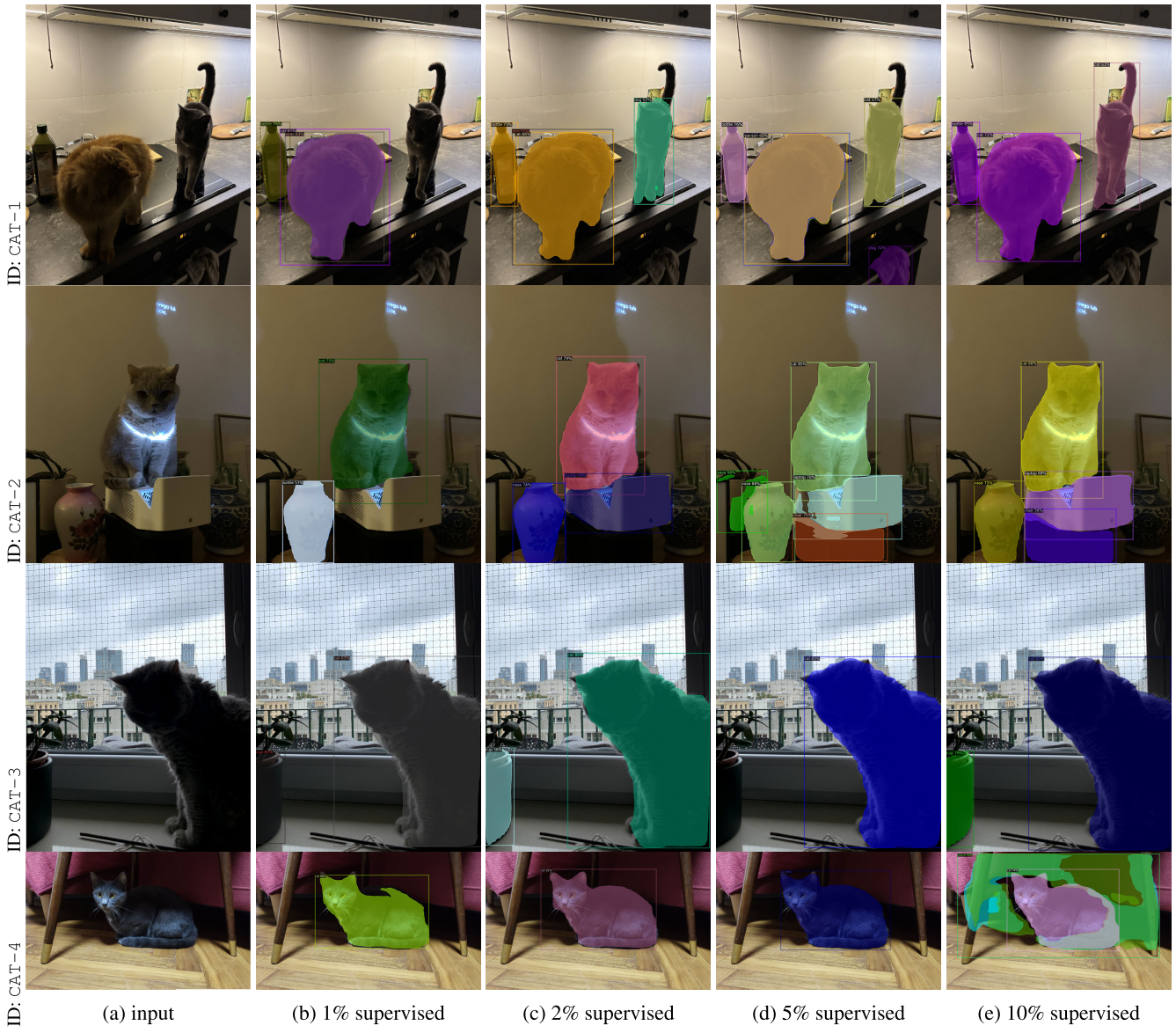

FIGURE 6. Qualitative Polite Teacher results with different supervision regimes on uneven light conditions, tested on cats of the corresponding author.

of validation on other datasets. Similarly, more single-stage detectors, as well as two-stage detectors can be taken into consideration. For more robust evaluation results, several runs of the experiments to explore variance on different values of supervision might be carried out. Therefore, future work should consider validating methods with more detectors, backbones and datasets. Providing a direct comparison with Noisy Boundaries [26] might be considered as well. A natural next step would consider taking pseudo-bounding boxes and pseudo-centre-ness regression into account.

ACKNOWLEDGMENT

Some experiments were performed using the Entropy cluster at the Institute of Informatics, University of Warsaw.

AUTHOR CONTRIBUTIONS

Dominik Filipiak: (50% of the work) conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, writing – review & editing, visualization, supervision, project administration. **Andrzej Zapała:** (20% of the work) software, validation, formal analysis, investigation, data curation, writing – review & editing, visualization. **Piotr Tempczyk:** (10% of the work) writing – review & editing, supervision. **Anna Fensel:** (5% of the work) resources, writing – review & editing, funding acquisition. **Marek Cygan:** (15% of the work) conceptualization, methodology, resources, writing – review & editing, supervision, funding acquisition.

REFERENCES

[1] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–8.

[2] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[3] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda, "Unbiased teacher for semi-supervised object detection," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[4] Y.-C. Liu, C.-Y. Ma, and Z. Kira, "Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9809–9818.

[5] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6402–6411.

[6] T. Berrada, C. Couprie, K. Alahari, and J. Verbeek, "Guided distillation for semi-supervised instance segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Jan. 2024, pp. 475–483.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–11.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[11] Y. Lee and J. Park, "CenterMask: Real-time anchor-free instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13903–13912.

[12] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[14] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[15] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3041–3050.

[16] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19358–19369.

[17] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[18] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6748–6758.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[20] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–12.

[21] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[22] A. Peláez-Vegas, P. Mesejo, and J. Luengo, "A survey on semi-supervised semantic segmentation," 2023, *arXiv:2302.09899*.

[23] I. J. Goodfellow, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[24] Q. H. Le, K. Youcef-Toumi, D. Tsetserukou, and A. Jahanian, "GAN mask R-CNN: Instance semantic segmentation benefits from generativeadversarial networks," 2020, *arXiv:2010.13757*.

[25] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," in *Proc. Int. Conf. Learn. Represent.*, 2022.

[26] Z. Wang, Y. Li, and S. Wang, "Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16805–16814.

[27] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Ross Girshick, Georgia, Jun. 2018, pp. 4119–4128.

[28] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.

[29] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, "A simple semi-supervised learning framework for object detection," 2020, *arXiv:2005.04757*.

[30] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 516–520.

[31] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. (2019). *Detectron2*. [Online]. Available: https://github.com/facebookresearch/detectron2

[32] T. Lin, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

DOMINIK FILIPIAK received the M.Eng. degree in computing from the Faculty of Computing, Poznań University of Technology, and the Ph.D. degree in economics from the Department of Information Systems, Poznań University of Economics and Business. He is currently pursuing the Ph.D. degree in computer science with Universität Innsbruck. In his scientific work, he explores modern methods of artificial intelligence, with a special focus on computer vision and imperfect data conditions. He has been with AI, since 2011. He is also the CEO and the Co-Founder of the Polish Branch of Perelyn, an advanced AI consulting company. He teaches with the Institute of Computer Science, Faculty of Mathematics, Computer Science, and Mechanics, University of Warsaw.

ANDRZEJ ZAPAŁA received the B.Sc. degree from the Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, where he is currently pursuing the M.Sc. degree. In his work, he tackles current challenges in modern artificial intelligence and machine learning.

PIOTR TEMPCZYK received the degree from the Faculty of Physics, University of Warsaw. He is currently pursuing the Ph.D. degree with the Faculty of Mathematics, Computer Science, and Mechanics, University of Warsaw. He has been involved in data analysis, machine learning, and artificial intelligence, since 2012. He also works in industry, in the field of ML. He conducts research with NASK and leads the PL4AI Research Group.

ANNA FENSEL is Full Professor in Artificial Intelligence and Data Science at Wageningen University, Wageningen, the Netherlands. Before she was Associate Professor at Wageningen University and at STI Innsbruck, Department of Computer Science, University of Innsbruck, Austria. Earlier she worked as a Senior Researcher at FTW-Telecommunications Research Center Vienna, Austria, and a Research Fellowat the University of Surrey, UK, and as a project employee at DERI Innsbruck, University of Innsbruck, Austria. She has earned both her habilitation and her doctoral degree in Computer Science at the University of Innsbruck in 2018 and 2006, respectively. Prior to that, she has received a diploma in Mathematics and Computer Science equivalent to the Master degree from Novosibirsk State University, Russia, in 2003.

MAREK CYGAN is currently a Professor with the Institute of Informatics, University of Warsaw. He has a background in competitive programming (ACM ICPC and Google Code Jam), for a few years, he did research in algorithms but recently switched to machine learning in robotics. He is leading the Robot Learning Group, University of Warsaw.