2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 979-8-3503-6547-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPRW63382.2024.00552

Knowledge Distillation for Efficient Instance Semantic Segmentation with Transformers

Maohui Li[1] Michael Halstead[1]
Chris McCool[1,2]
[1]University of Bonn, [2]Lamarr Institute for Machine Learning and Artificial Intelligence
{mlii, michael.halstead, cmccool}@uni-bonn.de

Abstract

Instance-based semantic segmentation provides detailed per-pixel scene understanding information crucial for both computer vision and robotics applications. However, state-of-the-art approaches such as Mask2Former are computationally expensive and reducing this computational burden while maintaining high accuracy remains challenging. Knowledge distillation has been regarded as a potential way to compress neural networks, but to date limited work has explored how to apply this to distill information from the output queries of a model such as Mask2Former.

In this paper, we match the output queries of the student and teacher models to enable a query-based knowledge distillation scheme. We independently match the teacher and the student to the groundtruth and use this to define the teacher to student relationship for knowledge distillation. Using this approach we show that it is possible to perform knowledge distillation where the student models can have a lower number of queries and the backbone can be changed from a Transformer architecture to a convolutional neural network architecture. Experiments on two challenging agricultural datasets, sweet pepper (BUP20) and sugar beet (SB20), and Cityscapes demonstrate the efficacy of our approach. Across the three datasets the student models obtain an average absolute performance improvement in AP of 1.8 *and* 1.9 *points for ResNet-50 and Swin-Tiny backbone respectively. To the best of our knowledge, this is the first work to propose knowledge distillation schemes for instance semantic segmentation with transformer-based models.*

Index Terms–Computer Vision for Agriculture Automation, Knowledge Distillation, Efficient Instance Segmentation, Transformer

1. Introduction

The large application of computer vision algorithms, such as detection, semantic segmentation, and instance-based semantic segmentation brings the potential to greatly improve efficiency to agricultural automation. Automatic fruit picking, weed removal, and pesticide drip irrigation have been made possible by the introduction of dense predictions [1]. Instance segmentation provides a wealth of information like per-pixel classification and instance location of objects, which contributes to agricultural efficiency and makes it possible to convert agricultural production from human labor to automation. Despite these advances, it still remains highly challenging to deploy vision algorithms in agriculture since it is vulnerable to clutter from leaves and other crop, as well as highly variable lighting conditions. Furthermore, the limited computational and energy resources on edge devices constrains deployment on robots. Therefore, efficient and accurate models capable of instance-based segmentation are integral to enable real-world deployment.

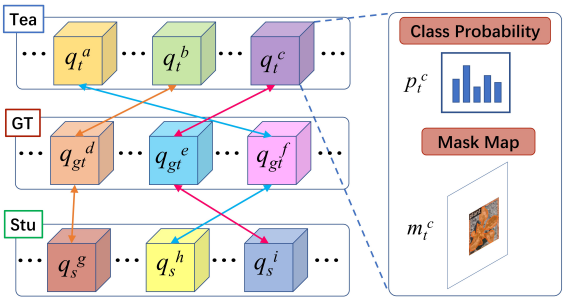

Figure 1. A visualization of our bipartite query-based matching between the *teacher* and the *student*. We compare first association of the predicted queries of the model, either *teacher* or *student*, to the groundtruth to obtain *teacher-gt* and *student-gt* respectively. We then use this to find the association of the predicted queries for the *teacher-student*.

Transformer-based models were introduced to computer vision to further improve accuracy after the successful application of vanilla transformers [2] in natural language processing (NLP). DETR [3] was one of the first object detection models based on transformers, and it introduces the conception of object query which only contain features from one instance. More recently, the Mask2Former [4]

2160-7516/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPRW63382.2024.00552

architecture was proposed which is capable of state-of-the-art semantic, instance-based, and panoptic segmentation by extracting masked attention within predicted mask regions. Despite the improved accuracy provided by transformer-based models their complicated structures inevitably increases the computational complexity and hampers their application in real-time scenarios.

One of the most effective ways to improve the efficiency of deep neural networks is through knowledge distillation, by imparting knowledge from complex networks (*teachers*) to efficient networks (*students*). There are many kinds of knowledge that can be used to distill information from using features from the final or intermediate layers through to using the mutual relationship between features. However, the majority of these distillation schemes are designed for use with convolutional neural network (CNN) -based structures.

This means that distillation of joint image- and pixel-level classification of transformer-based networks are yet to be explored. Thus, how to distill instance-level knowledge from a transformer-based structure is the key contribution of this paper.

In this paper, we compress complex transformer-based models into more efficient networks while retaining a high degree of accuracy. First, we produce an optimal bipartite matching scheme between the queries from the *teacher* and the *student* as shown in Figure 1 by using the Hungarian algorithm [3]. Second, we train our efficient networks by distilling the class probabilities and mask maps from the *teacher* network. For our experiments we perform distillation using the Mask2Former [4] architecture due to its high accuracy for instance-based semantic segmentation and the ability to easily switch the backbone network. Finally, We show the validity of our approach by performing knowledge distillation in multiple domains: arable farmland, horticulture, and traffic scenarios.

Our results, on challenging agricultural and traffic datasets, demonstrate that our knowledge distillation scheme can be employed to learn efficient and accurate lightweight models. Our models are 2.3 or 2.0 times faster than the original complex *teacher* while only degrading AP (average precision) performance by as little as 2.8 or 4.0 points, and in one case outperforms the *teacher* by 1.0 point. The main contribution of our paper is that we establish pair-wise query matching between transformer-based models with different complexities (backbones and number of queries) and distill the query-based knowledge from the *teacher* to the *student*.

2. Related Work

2.1. Instance Segmentation

Instance-based semantic segmentation can be regarded as the process of labeling pixels with categories and object ids. This pixel-wise classification is fundamental for many advanced computer vision based tasks, such as medical images analysis [5], autonomous driving [6], and scene understanding [7]. The methods can be roughly divided into two main categories: single-stage and two-stage methods. Two-stage methods dominate instance segmentation and can be further divided into top-down methods and bottom-up methods. Top-down methods [8–10] predict bounding-boxes first and then generate the instance masks within these boxes which means that the final performance is highly dependent on detection results. In contrast, bottom-up methods [11, 12] classify pixels into the corresponding categories and from this post-processing is applied to form the instances which means that the final performance is highly dependent on the post-processing approach. Single-stage methods jointly perform detection and segmentation, and are further divided into anchor-based and anchor-free methods. For anchor-based methods [13, 14], most detectors rely heavily on pre-defined anchors which vary in scale depending on the target dataset. These techniques also rely on handcrafted approaches like non-maximal suppression (NMS) to remove redundancies, which can increase the computational burden. For anchor-free methods [15, 16], they distinguish instances on the basis of predicted locations and shapes of the objects. A limitation of anchor-free methods is that their performance generally decreases when many instances overlap, this is because each grid can only predict one location and mask.

2.2. Transformer-Based Dense Prediction

A variety of transformers have been designed for vision tasks [17, 18]. A standard backbone for various dense vision prediction tasks is the Swin-Transformer [19] which consists of a hierarchical architecture with multi-scale feature maps. Inspired by the transformer structures, recent work has exploited self-attention to capture the long-range relationships needed to perform instance-based semantic segmentation. A query-based instance segmentation method based on the structure of Sparse R-CNN [20] is proposed by Fang et al. [21]. In the approach, a mask pooling operator is designed to extract the current stage instance mask features and a dynamic convolution module is employed to set a linkage between mask features and query embedding. In ISTR [22], three task-specific heads and a fixed mask decoder work together to accomplish the classification, localization, and segmentation prediction by taking in the image-level features and learnable position embedding. Dong et al. [23] propose a structure to complete clas-

sification, bounding box regression, and mask segmentation in one head linking segmentation and detection as a unified query representation. However, complex transformer-based models do not currently meet the real-time requirements of robotic platforms.

2.3. Knowledge Distillation

Knowledge distillation has proven effective at compressing models while maintaining performance. This enables a lightweight network (*student*) to boost its performance learning soft labels from a more complicated (*teacher*) network [24]. The form of knowledge is traditionally divided into three categories when performing distillation with CNN-based models: response-based knowledge, feature-based knowledge, and relation-based knowledge respectively [25]. Response-based knowledge relies on the response of the last output layer of the used networks [26, 27]. Feature-based knowledge uses the features from selected intermediate layers, and an adaptation layer is often required to address the dimension mismatch between the *teacher* and the *student* when doing distillation [28, 29]. Relation-based knowledge focuses more on the relationships between features from different network layers or data samples, with typical examples being [30].

These distillation techniques have been shown to perform well for CNN-based networks, however, for transformer-based networks they are not directly applicable due to the inherently different network structures. Due to these differences self-attention based knowledge distillation approaches are required. Lin et al. [31] propose the target-aware transformer. Their approach assumes that the semantic information usually varies on the same spatial location since the *teacher*'s and the *student*'s receptive field are different. The model generates the similarity between each pixel of the *teacher*'s feature and all spatial locations of the *student*'s features during the distillation process. The performance of their technique surpasses the state-of-the-art methods by a significant margin on common benchmarks. In [32], the authors find that the dominant factor that affects the distillation performance lies in inductive *teacher* bias rather than *teacher* accuracy. The *student* is more likely to learn diverse knowledge by transferring the inductive biases from both the CNN and involution-based neural network (INN) *teacher* when distilled. Touvron et al. [33] propose a novel attention distillation mechanism and achieve state-of-the-art performance on ImageNet. The authors exploit a distillation token, similar to the classification token within ViT [34], to enable the *student* to learn the attention maps from the *teacher*. To gain cross-architecture knowledge, a novel distillation scheme is proposed by Liu et al. [35] to combat the heterogeneous architectures' gap. Two projectors, one for partial cross-attention and one for group-wise linear, are designed to align the *student*'s immediate features into the *teacher*'s attention space during this process. As the complex structure of the *teacher* consumes a large amount of resources during the process of supervising the *student*, a framework [36] that stores the *teacher*'s predictions in advance was proposed.

The distillation methods with transformed-based models mentioned above focus mainly on image-level or pixel-level classification, but new distillation schemes on instance-based semantic segmentation and other dense prediction paradigms are currently not explored. In this paper we investigate this dense prediction by using a bipartite matching scheme to distill information from a *teacher* to a *student*.

3. Proposed Approach

In this paper, we propose a knowledge distillation scheme for instance-based semantic segmentation with transformer-based models. We build an ordered permutation for queries from the *student* and the *teacher* in an innovative way, then we distill the query-based knowledge based on the established matching. Finally, we test our scheme on two agricultural datasets and a common city scene dataset. The approach consists of three aspects.

1. We adopt Mask2Former as the framework for both *teacher* and *student* networks. For the *teacher* network we use the Swin-Large backbone [19] with a high number of queries. For the *student* networks we explore the use of both ResNet [37] and Swin for the backbones with fewer queries to explore a trade-off between efficiency and accuracy.
2. We build a new matching technique to facilitate the query-based distillation process. We match the unordered queries from a set of predictions from the *teacher* and the *student* by calculating their instance similarity using class probabilities and mask maps. We achieve this by exploiting the available groundtruth labels and the matching scheme is explicitly described in III-B.
3. Based on the result of the matching scheme, we perform instance-based knowledge distillation using both the *teacher* and groundtruth labels. For this, we define a *teacher* to *student* loss function that incorporates both class probabilities and segmentation masks. When using both the *teacher* and groundtruth labels we combine them using a weighted loss function. This is described in III-C.

3.1. Teacher and Student Network Architectures

We adopt Mask2Former as the framework, which is an encoder-decoder structure with self-attention layers. Mask2Former is composed of a backbone, pixel decoder, and transformer decoder which results in N queries which are used to predict the class probabilities (including object or not) and the associated per-pixel mask. The back-

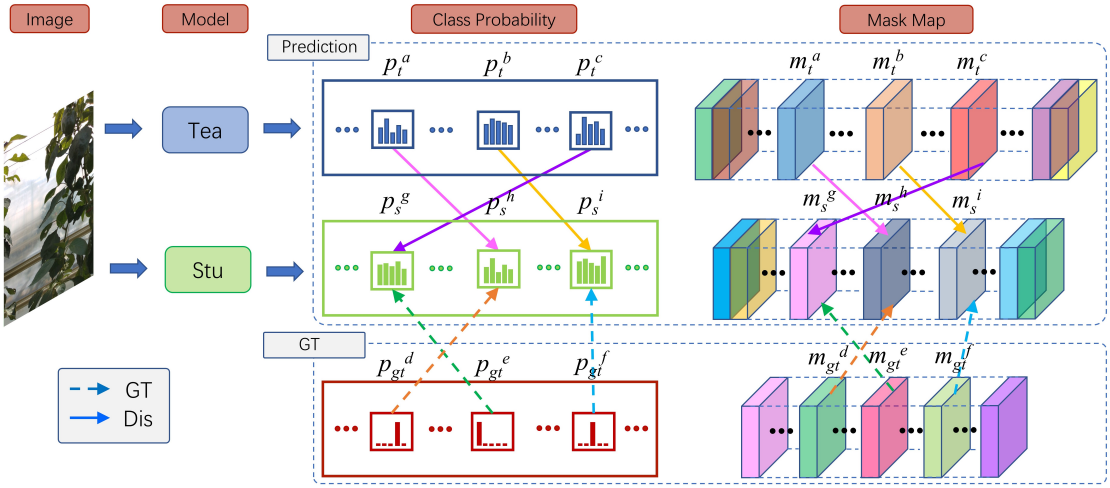

Figure 2. An illustration of our instance knowledge distillation process. Given the association of the predicted queries from the *teacher* to the *student* model we then use an L_2 loss to perform joint distillation on the class probabilities and mask predictions. In this process, only queries with effective instances will be distilled and other redundant ones are discarded.

bone extracts information from an image and outputs features with different resolutions. The pixel decoder generates high-resolution per-pixel embedding by sampling the features from the previous backbone. The transformer decoder processes the long-range relationship of the image information and generates object queries, with which the final heads predict the probabilities and the masks. More details on this approach can be found in [4].

We take competitively performed and commonly used ResNet and Swin-Transformer as the backbones of our models. The model inference complexity can be varied by selecting a set of backbones, from ResNet-18 to ResNet-101 or from Swin-T (tiny version) to Swin-L (large version). We set the *teacher* backbone to be the top performer (Swin-L) and the *student* uses either ResNet-50 or Swin-T as backbones as they computationally more efficient. The *teacher* backbone was selected to ensure that it can capture deeper relationships within the scene and it consists of 200 queries. As the *student* backbone is simpler we reduce its number of queries to be 100.

3.2. Student-Teacher Query Matching

In order to perform knowledge distillation we need to find matching queries between the *teacher* and *student* models. As shown in Figure 1, we establish the *teacher-student* matching by exploiting the existing groundtruth labels (*gt*). We do this by first resolving the *teacher-gt* and *student-gt* associations. From this, we can then derive the *teacher-student* associations, $\hat{\phi}$.

To find the optimal bipartite matching, *teacher-gt* and *student-gt* we exploit the commonly used Hungarian algorithm in previous work [3]. This establishes the query permutation by optimizing the total cost involved by combining both class probabilities and instance mask predictions. For this we use the following matching loss (cost),

$$\mathcal{L}_{match} = \delta_{cls}\mathcal{L}_{cls} + \delta_{msk}\mathcal{L}_{msk}, \tag{1}$$

where $\delta_{cls} = 2$ and $\delta_{msk} = 5$. $\mathcal{L}_{cls}$ and $\mathcal{L}_{msk}$ are the corresponding losses of class probabilities and mask predictions.

For the matching cost of classes, we only use probabilities from the query set that match to an object (i.e. $C_i \neq \emptyset$),

$$\mathcal{L}_{cls} = -\mathbb{1}_{C_i \neq \emptyset}\hat{p}_{\sigma(i)}(C_i), \tag{2}$$

where $\hat{p}_{\sigma(i)}$ is the predicted probability of the class, and C_i is the groundtruth class label.

As for the matching cost of masks, we calculate similarities of pixels and overlaps between instances,

$$\mathcal{L}_{msk} = -\mathbb{1}_{C_i \neq \emptyset}[\mathcal{L}_{dic}(m_i, \hat{m}_{\sigma(i)}) + \mathcal{L}_{xe}(m_i, \hat{m}_{\sigma(i)})] \tag{3}$$

where m_i and $\hat{m}_{\sigma(i)}$ are masks from *gt* and predictions in one image, and $\mathcal{L}_{dic}$ and $\mathcal{L}_{xe}$ are the Dice loss and Cross Entropy loss respectively.

3.3. Knowledge Distillation Loss function

As shown in Figure 2, the *student* uses both the *teacher* information and groundtruth labels during distillation. This leads to our loss function being composed of two parts, the groundtruth loss and the *teacher-student* loss,

$$\mathcal{L}_{all} = \mathcal{L}_{gt} + \alpha\mathcal{L}_{dis}, \tag{4}$$

where α is the balancing weight which varies according to the different datasets. $\mathcal{L}_{gt}$ and $\mathcal{L}_{dis}$ are the corresponding losses from the hard labels and soft labels during distillation.

For the groundtruth loss, we compare the similarity of the predicted class possibilities and mask predictions with the corresponding hard labels as,

$$\mathcal{L}_{gt} = \delta_{cls}\mathcal{L}_{xe}(GT_{cls}, S_{cls}) + \delta_{msk}\mathcal{L}_{msk} \quad (5)$$

$$\mathcal{L}_{msk} = \mathcal{L}_{dic}(GT_{msk}, S_{msk}) + \mathcal{L}_{xe}(GT_{msk}, S_{msk}), \quad (6)$$

where GT_{msk} is the groundtruth mask, S_{msk} is the predicted mask, GT_{cls} is the groundtruth class, and S_{cls} is the predicted class. The balancing weights are set to $\delta_{cls} = 2$ and $\delta_{msk} = 5$. The class-based loss $\mathcal{L}_{xe}$ is a standard cross-entropy loss. The mask-based loss $\mathcal{L}_{msk}$ uses a weighted combination of a standard cross entropy loss and a Dice loss $\mathcal{L}_{dic}$; similar to [4] we use equal weights of 5 for each.

The distillation loss is defined as follows,

$$\mathcal{L}_{dis} = \sum_{i}^{N} ||T^i_{cls}, S^{\hat{\phi}(i)}_{cls}||^2_2 + \beta \sum_{i}^{N} ||T^i_{msk}, S^{\hat{\phi}(i)}_{msk}||^2_2 \quad (7)$$

where T^i_{cls} is the *teacher* class logits, $S^{\hat{\phi}(i)}_{cls}$ is the *student* class logits, T^i_{msk} is the *teacher* predicted mask logits, $S^{\hat{\phi}(i)}_{msk}$ is the *student* predicted mask logits, and β is the balancing weight. The i-th *teacher-student* association is denoted as $\hat{\phi}(i)$. To compare the *teacher* and *student* logits we use an L_2 loss. To address the imbalance between the class and mask losses we set the balancing weight $\beta = 0.02$.

4. Experimental Setup and Results

We evaluate our instance semantic segmentation systems on two challenging agricultural datasets, BUP20 and SB20, and Cityscapes dataset. BUP20 [38] is a sweet pepper dataset (glasshouse) which consists of 280 images which are split into three sets with 124 images to train, 63 images to validate and 93 images to evaluate (test). SB20 [39] is a sugar beet dataset (arable farming) which consists of 143 images which are split into three sets with 71 images to train, 37 images to validate and 35 images to evaluate (test). Cityscapes [40] is a traffic scene dataset which consists of 5000 images, which are split into three sets with 2975 to train, 500 to valid and 1525 to evaluate (test). Samples of the three real-world datasets can be seen in Figure 3. BUP20 and Cityscapes both have high levels of occlusion while SB20 has large variation due to different growth stages of the plants.

Table 1. Instance knowledge distillation results on BUP20 in terms of AP, $AP50$ and $AP75$.

models	AP	$AP50$	$AP75$
teacher ($M2F_{sl}$)	51.3±0.1	81.1±0.1	53.3±0.3
baseline ($M2F_{r50}$)	44.8±0.3	72.3±0.8	46.2±0.6
dis (S_{r50})	**46.1±0.8**	**73.4±0.6**	**48.3±0.9**
baseline ($M2F_{st}$)	47.7±0.3	76.9±0.2	49.6±0.3
dis (S_{st})	**48.5±0.3**	**77.2±0.4**	**50.9±0.7**

All of our models are implemented in Detectron2 [41] and trained on an NVidia A6000 GPU. We fine-tune all the models with weights pre-trained on the COCO dataset [42] and do this 3 times to get the mean and variance of the performance. The only exception is that on Cityscapes we do not train multiple *teacher* models but instead use the pre-trained model provided by Mask2Former. For BUP20 and SB20, we set AdamW optimizer [43] with a step learning rate of $\gamma = 1e^{-4}$ and $\gamma = 1e^{-5}$ respectively for the backbone ResNet and Swin-Transformer with a batch $b = 1$; the small batch size is due to the low number of training images. We search for the optimal weight α to combine groundtruth and *teacher* labels using ranges 0.2 to 5.0. For Cityscapes, we set AdamW optimizer with a step learning rate of $\gamma = 1e^{-4}$ with a batch $b = 16$. Here we set the search for α to in the range 0.2 to 2.0.

We report performance primarily using average precision (AP). Our *teacher* network has 200 queries with a Swin-Large backbone, referred to as $M2F_{sl}$. The *student* networks have only 100 queries and use either a ResNet-50 (S_{r50}) or Swin-Tiny (S_{st}) backbone.

4.1. Experiments on Agricultural Data

4.1.1 Results on BUP20

Table 1 outlines the results for BUP20 where there is a clear improvement based on our distillation scheme. Our most efficient network, S_{r50}, obtains the greatest performance boost with an absolute AP improvement of 1.3 points when compared to its direct baseline $M2F_{r50}$ (44.8 to 46.1). The other distilled network, S_{st}, also improves results with an absolute improvement in AP of 0.8 points.

We attribute the greater increase of the S_{r50} model to the overall performance gap between the baseline ($M2F_{r50}$) and the *teacher* network when compared to that of the $M2F_{st}$ to the *teacher*. The difference between the two baseline approaches ($M2F_{r50}$ and $M2F_{st}$) is believed to be based on more informative features output by the Swin-T backbone. Interestingly, it can be seen that the main performance gain for the distilled models occurs at higher APs.

It can be seen that there are large performance gains for AP75 but much lower gains for AP50. The performance gain for AP75 is 2.1 and 1.3 points for S_{r50} and S_{st} respectively. By comparison the performance for AP50 is 1.1 and 0.3 points for S_{r50} and S_{st} respectively. This indicates that the distillation approach is providing models which provide considerably more accurate semantic segmentation masks which is important for downstream tasks for automating the estimation of phenotypic attributes (e.g. size of fruit) as well as robotic tasks such as harvesting.

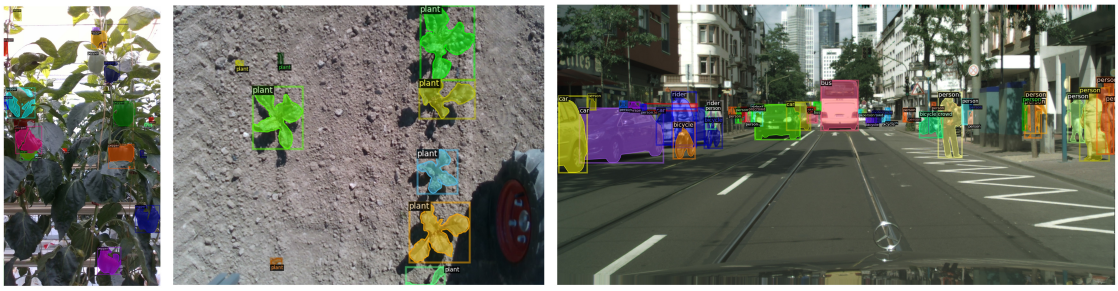

Figure 3. Example images from the three datasets with groundtruth instance segmentation masks of BUP20, SB20 , and Cityscape from left to right.

4.1.2 Results on SB20

The results for SB20, see Table 2, demonstrate that our distillation approach consistently improves the performance of our models. For SB20, the *student* network S_{st} is able to achieve the greatest performance improvement. For the S_{r50} network we achieve an absolute improvement in AP of 2.0 points (34.9 to 36.9) compared to 3.0 points for the more complex S_{st} (36.4 to 39.4). The impact of our distillation scheme is most evident when comparing S_{st} to the *teacher* network where we improve our performance by 1.0 point, however, we note that this improvement is still within the bounds of the variance of the models where the *teacher* network has an AP of 38.4±0.7 and the distilled model S_{st} an AP of 39.4±0.5.

Similar to BUP20 the biggest performance improvements occur for higher APs which can be seen by examining the performance gains of AP50 vs AP75. The performance gain for AP75 is 3.5 and 5.2 points for S_{r50} and S_{st} respectively. By comparison the performance for AP50 is 1.1 and 1.5 points for S_{r50} and S_{st} respectively. For arable farming data such as SB20, providing considerably more accurate segmentation masks is important for phenotyping tasks such as estimating leaf-area as well as precise robotic weeding.

4.2. Ablation on Cityscapes

To further validate the generalizability ability of our distillation system, we apply this to Cityscapes. Cityscapes is very different to BUP20 and SB20 as it consists of pedestrians and vehicles. For our analysis we use a pre-trained *teacher* model by downloading the online pre-trained weights from Mask2Former [4].

Our results in Table 3 demonstrate our distillation approach is also effective on this data. The *student* S_{r50} gains an absolute performance increase in AP of 2.1 points from 35.8 to 37.9, and the *student* S_{st} gains 1.8 points improvement from 37.9 to 39.7. Despite these considerable performance gains there is still a large gap in performance between the *students* and the *teacher*. For the S_{r50} network the absolute performance in terms of AP is 5.8 points while for S_{st} it is 4.0 points, when compared to the *teacher*.

4.3. Best Student Model and Inference Time

In the previous experiments we demonstrate that our distillation approach consistently improves performance. Overall, we demonstrate that our distillation scheme achieves considerable improvements with an average absolute performance improvement in terms of AP of 1.8 points for S_{r50} and 1.9 points for S_{st} across the three datasets. On average, over the three datasets, the two *student* models are 2.0 and 2.3 times faster, for S_{st} and S_{r50} respectively. Furthermore, the S_{st} model consistently outperforms S_{r50} and so in most cases it would be the preferred distilled model. However, for SB20 the relative performance degradation is smallest, with S_{r50} dropping by 1.5 points for AP which is a 3.9% relative reduction in performance. Therefore, for this case it might be considered as preferable if the lower inference time is considered imperative as it is 14 milliseconds, or 14%, faster than S_{st}; SB20 has a smaller image size so that relative speed difference is lower than on other high resolution datasets.

Table 2. Instance knowledge distillation results on SB20 in terms of AP, $AP50$ and $AP75$.

models	AP	$AP50$	$AP75$
teacher ($M2F_{sl}$)	38.4±0.7	79.2±0.9	31.9±0.7
baseline ($M2F_{r50}$)	34.9±1.3	75.3±2.0	28.1±2.1
dis (S_{r50})	**36.9±0.5**	**76.4±0.9**	**31.6±0.9**
baseline ($M2F_{st}$)	36.4±0.8	78.6±0.6	29.8±1.5
dis (S_{st})	**39.4±0.5**	**80.1±0.5**	**35.0±1.0**

Table 3. Instance knowledge distillation results on Cityscapes in terms of AP, and $AP50$.

models	AP	$AP50$
teacher ($M2F_{sl}$)	43.7	71.3
baseline ($M2F_{r50}$)	35.8±1.0	62.6±1.1
dis (S_{r50})	**37.9±0.3**	**64.2±0.3**
baseline ($M2F_{st}$)	37.9±0.8	65.1±0.6
dis (S_{st})	**39.7±0.7**	**66.4±1.4**

5. Summary

In this paper, we have proposed a bipartite matching approach to perform knowledge distillation on output queries from Mask2Former, transformer-based network. The bipartite matching allows us to associate the queries predicted by the *student* and the *teacher*. Using this association we then distill the corresponding query-based class probabilities and instance masks from the *teacher* to the *student*. We apply this to the *student* models with vastly different backbones which consist either of a transformer backbone (Swin-Tiny) and even a DCNN backbone (ResNet-50). To be the best of our knowledge, this is the first that such an approach for knowledge distillation has been proposed.

We evaluate our knowledge distillation scheme on two challenging agricultural datasets as well as Cityscapes which consists of pedestrian and vehicle data. In all cases, applying our approach leads to improved performance for the distilled models with an average absolute performance improvement in terms of AP of 1.8 points for S_{r50} and 1.9 points for S_{st} across the three datasets. In particular, our approach leads to more precise detections as demonstrated by higher values for AP75 than AP50. Overall, we show that simple *student* networks trained with our instance knowledge distillation scheme can retain a high accuracy with faster inference than the *teacher* model. Future work should examine the potential implication of changing not just the backbone but also the pixel decoder to further improve the tradeoff between accuracy and computational efficient.

References

[1] M. Halstead, P. Zimmer, and C. McCool, "A cross-domain challenge with panoptic segmentation in agriculture," *The International Journal of Robotics Research*, p. 02783649241227448, 2024. 1

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 1

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229. 1, 2, 4

[4] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299. 1, 2, 4, 5, 6

[5] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Processing*, vol. 16, no. 5, pp. 1243–1267, 2022. 2

[6] L. Guan and X. Yuan, "Instance segmentation model evaluation and rapid deployment for autonomous driving using domain differences," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4050–4059, 2023. 2

[7] I. Balazevic, D. Steiner, N. Parthasarathy, R. Arandjelović, and O. Henaff, "Towards in-context scene understanding," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 2

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969. 2

[9] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6409–6418.

[10] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F.-Y. Wang, "Mask ssd: An effective single-stage approach to object instance segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2078–2093, 2019. 2

[11] D. Neven, B. D. Brabandere, M. Proesmans, and L. V. Gool, "Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 8837–8845. 2

[12] N. Gao, Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, and K. Huang, "Ssap: Single-shot instance segmentation with affinity pyramid," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 642–651. 2

[13] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166. 2

[14] X. Chen, R. Girshick, K. He, and P. Dollár, "Tensormask: A foundation for dense object segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2061–2069. 2

[15] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 649–665. 2

[16] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17 721–17 732, 2020. 2

[17] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 124–12 134. 2

[18] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "Mpvit: Multi-path vision transformer for dense prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7287–7296. 2

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 2, 3

[20] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 454–14 463. 2

[21] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6910–6919. 2

[22] J. Hu, L. Cao, Y. Lu, S. Zhang, Y. Wang, K. Li, F. Huang, L. Shao, and R. Ji, "Istr: End-to-end instance segmentation with transformers," *Energy*, vol. 50, p. 100. 2

[23] B. Dong, F. Zeng, T. Wang, X. Zhang, and Y. Wei, "Solq: Segmenting objects by learning queries," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 898–21 909, 2021. 2

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 3

[25] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021. 3

[26] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5311–5320. 3

[27] H. Bai, H. Mao, and D. Nair, "Dynamically pruning segformer for efficient semantic segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3298–3302. 3

[28] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 319–12 328. 3

[29] Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, and C. Yuan, "Masked generative distillation," in *European Conference on Computer Vision*. Springer, 2022, pp. 53–69. 3

[30] S. An, Q. Liao, Z. Lu, and J.-H. Xue, "Efficient semantic segmentation via self-attention and self-distillation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15 256–15 266, 2022. 3

[31] S. Lin, H. Xie, B. Wang, K. Yu, X. Chang, X. Liang, and G. Wang, "Knowledge distillation via the target-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 915–10 924. 3

[32] S. Ren, Z. Gao, T. Hua, Z. Xue, Y. Tian, S. He, and H. Zhao, "Co-advise: Cross inductive bias distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 16 773–16 782. 3

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357. 3

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 3

[35] Y. Liu, J. Cao, B. Li, W. Hu, J. Ding, and L. Li, "Cross-architecture knowledge distillation," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 3396–3411. 3

[36] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 68–85. 3

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 3

[38] C. Smitt, M. Halstead, T. Zaenker, M. Bennewitz, and C. McCool, "Pathobot: A robot for glasshouse crop phenotyping and intervention," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2324–2330. 5

[39] A. Ahmadi, M. Halstead, and C. McCool, "Bonnbot-i: A precise weed management and crop monitoring platform," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 9202–9209. 5

[40] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[41] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019. 5

[42] T.-Y. 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755. 5

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2018. 5