Towards Efficient Instance Segmentation with Hierarchical Distillation

Ziwei Deng Quan Kong Tomokazu Murakami
Hitachi, Ltd. R&D Group, Japan
{ziwei.deng.qq, quan.kong.xz, tomokazu.murakami.xr}@hitachi.com

Abstract

Recently, instance segmentation models have been developed to a great promising accuracy on public benchmarks. However, these models are too heavy to be applied for real applications due to their low inference speed. In this paper, we propose a faster instance segmentation model utilizing a teacher-student learning framework that transfers the knowledge obtained by a well-trained teacher model to a lightweight student model. In addition to the conventional strategy of knowledge distillation in classification or semantic segmentation networks which are both single-task networks, we investigate a hierarchical distillation (H-Dis) framework for structure information distillation on multi-task learning based instance segmentation. H-Dis consists of two distillation schemes: representation distillation that distills pair-wise quantized feature maps shared by multi-heads, and semantic distillation that makes sure to distill each head information in an instance level. In particular, we present channel-wise distillation for the segmentation head to achieve instance-level mask knowledge transfer. To evaluate our approach, we carry out experiments with different settings of distillation methods on different datasets Pascal VOC and Cityscapes. Our experiments prove that our approach is effective for accelerating instance segmentation models with less accuracy drop under limited computing resources.

1. Introduction

In recent years, the performance of instance segmentation has been dramatically improved by deep convolutional neural networks, which has made an attractive possibility to apply it to real-world applications, such as surveillance systems, auto-driving systems and medical systems. In general, a stronger DNN model has a deeper and heavier network design which makes it suffer from low inference speed, while on the contrary, the speed is a key requirement in most real-world applications. Thus, a faster and lighter instance segmentation model with promising accuracy is highly demanded.

Previous researchers discovered some possible ways to accelerate the DNN models. Model compression [7, 14, 27, 28] decomposes the weights in each layer to remove redundancy, recovering some accuracy by layer-wise reconstructions and fine-tuning. Model pruning [1, 24, 16] selects layer channels with higher importance or under sparse constraint. These kinds of methods achieve significant speed-up, but the accuracy drop is obvious as well, especially for complex tasks such as object detection and segmentation.

In order to enhance the accuracy of tiny or compressed models, knowledge transfer is a good way to transfer useful and effective knowledge learned by a cumbersome teacher model to a lightweight student model. Popular ways to transfer knowledge from one model to another are knowledge distillation[13, 26] and mimic learning[22, 17]. Conventional use of knowledge distillation has been widely proposed for classification networks [13, 15]. Meanwhile, for training compact semantic segmentation networks whose problem can be simply seen as pixel-wise classification, knowledge distillation can also be directly applied as pixel-wise distillation [25, 20]. However, both classification distillation and semantic segmentation distillation are single-task knowledge distillation. Applying knowledge distillation on popular multi-task instance segmentation networks is challenging because it is a structure information transfer with multiple kinds of semantic information in an instance level, that both classification and semantic segmentation are out scope of it.

In this paper, we propose a Hierarchical Distillation (H-Dis) framework for instance segmentation by taking both the distillation of outputs from middle layers as representation distillation and late layers as semantic distillation into consideration, to transfer the structure information from the cumbersome teacher network to the compact student network. H-Dis is a RoI(region of interest) based operation to optimize instance-level distillation. Representation distillation uses the quantized feature maps shared by each head network after RoI pooling as the representation information for distillation. Semantic distillation uses the outputs from each head network consisting of classification, bounding box regression and mask prediction. Different from pixel-

level classification with *softmax* in semantic segmentation, the mask branch in instance segmentation can be designed without inter-class competition (by a *sigmoid* instead of *softmax*), which gives large gains over *softmax* [11]. Thus, we present channel-wise distillation for the mask branch, in addition to the straightforward scheme by distilling the soft targets from classification networks and the bounding box coordinates from bounding box regression networks.

To summarize, the contributions of this paper are threefolds:

- We propose a novel hierarchical distillation framework for instance segmentation that utilizes both the middle and late layers' outputs from a teacher network as our distillation targets to train an efficient student network with a fast segmentation speed. To the best of our knowledge, this is the first trial for knowledge distillation on instance segmentation.
- We devise a new distillation loss for the mask branch in multi-task instance segmentation frameworks, with channel-wise distillation for teacher-student learning.
- The evaluations performed on large-scale benchmarks for demonstrating the effectiveness of the proposed method.

2. Related work

Instance segmentation: The task is to assign instance-level segmentation masks for each object in the image. Typically, the task requires segmentation, classification and box regression, which are accomplished separately or jointly relying on the region based methods, such as a Region Proposal Network (RPN). This paradigm includes most popular approaches, such as SDS [10], CFM[5] and MNC [6]. Most recently, FCIS [18] was proposed as a fully convolutional end-to-end solution for instance segmentation, performing position-sensitive output channels for mask estimation, classification and box regression jointly and simultaneously, which achieved both high accuracy and fast speed. Mask R-CNN [11] extended Faster R-CNN [9] by adding a FCN [21] branch for predicting an object mask in parallel with the class and box prediction branches. Mask R-CNN achieved competitive accuracy on public benchmarks, however the speed was not in their consideration. Considering both accuracy and speed, as well as the training time, we take the original FCIS as the baseline of this work.

Knowledge transfer: As a pioneering work of knowledge distillation, Hinton et al. [13] suggested a useful way to significantly improve the accuracy of a small model by transferring the generalization ability of an ensemble of networks, which leaded to a significant performance enhancement on the image classification task. The idea is to allow the student network to capture not only the information provided by the ground truth labels, but also the extra information about the finer structure learned by the teacher networks. Subsequent works tried to tackle the drawbacks of [13] by transferring intermediate features. Romero et al. [22] further developed this idea to make a thin and deep student network mimic the full feature maps of a wide and shallow teacher network. However, such assumptions are too strict since the capacities of teacher and student may differ greatly. In certain circumstances, it may adversely affect the performance and convergence. Chen et al. [2] proposed distillation losses for classification and box regression and applied mimic learning [22] in object detection networks. This work suggests a general way of applying knowledge transfer to multi-task networks. However, this work can only handle object detection problems and has no contribution in optimizing mimic learning for RoI-based tasks. Li et al. [17] extended mimic learning for object detection tasks, solving the above problem by transferring features merely in the RoIs. Xie et al. [25] investigated knowledge distillation for semantic segmentation networks, applying pixel-wise distillation on masks and consistency distillation which distills the information of mask boundaries. Liu et al. [20] further proposed a GAN(Generative adversarial network) based holistic distillation to match the masks generated by the teacher and the student. These works provide feasible ways for segmentation distillation, but are restricted to single-task semantic segmentation networks.

3. Method

We distill knowledge from a heavy and cumbersome teacher network to teach a light but efficient student network. Both of them follow the state-of-the-art instance segmentation architecture e.g., [18], with classification, box regression and mask branches in the heads of the network. The teacher network performs better than the student network in terms of segmentation accuracy, but with a slower processing speed. As [2] does, knowledge distillation is suggested to be added both in the head network and shared convolution layers.

Our method adopts both middle and late layers' output from the teacher network as our distillation targets to transfer the structure information to the student. The overview of our proposed method is shown in Figure.1.

3.1. Representation distillation

For instance segmentation, the feature maps extracted from shared convolutional layers will affect localization, classification and segmentation accuracy. It's essential to transfer from middle layers which contain numerous dark knowledge to facilitate the student model. Instance segmentation is a region-based work and the head networks are based on the pooled region proposals, so the region proposals play very important roles in this task. To distill

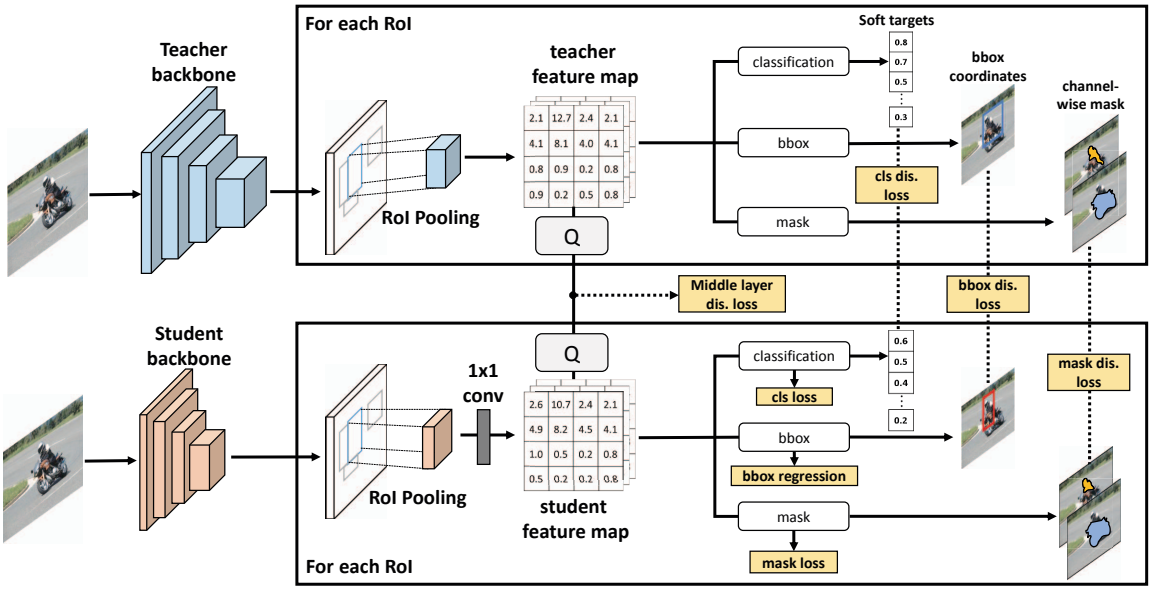

Figure 1. The overview of our proposed hierarchical distillation framework for instance segmentation. **Firstly**, we train the teacher network with all the training data and it only forwards the input image without back propagation in the distillation phase. **Secondly**, we operate representation distillation between RoIs on the feature maps of the teacher and student networks by L2 regularization as a middle layer distillation loss. Since most of the multi-task networks are heavier and deeper, a quantization operation [23] is adopted on the feature maps to make the student network converge faster and easier by learning from the teacher. **Thirdly**, we also distill the late layers' output as semantic information from the teacher but customized for each head, which consists of soft targets, bounding box coordinates and channel-wise masks for each RoI. Semantic distillation loss is composed of cross-entropy loss for classification distillation and L2 losses for box and mask distillation.

knowledge from middle layers, we follow [23] to introduce a quantization operation on the feature maps of the teacher and student networks. The quantization method is to discretize the output of both teacher network and student network in order to help the student to match the teacher better. The representation distillation loss function L_{rp_dis} is defined as:

$$L_{rp\_dis} = \frac{1}{2N} \sum_{i \in N} \|Q(f_t^i) - Q(r(f_s^i))\|_2^2 \qquad (1)$$

where Q is an element-wise quantization function that follows [23]. N is the number of proposals, and r is a function to transfer the student feature maps to the same size of the teacher's. Here, we use a 1×1 convolutional layer as r.

3.2. Semantic distillation

We adopt the outputs of the heads in the network which are kinds of semantic information. There are three branches used in the popular instance segmentation frameworks which consist of classification, box regression and mask prediction. Since the head networks are trained based on the region proposals generated by RPN, the region proposals for training the student network are the same as the teacher's.

Distillation for classification and box regression. Conventional use of knowledge distillation has been proposed for training classification networks, where categorization predictions of a teacher model are used as "soft targets" to guide the training of a student model. The neural networks typically produce class probabilities by using a *softmax* output layer that converts the classification score output z_i of the ith prediction computed for each class into a probability p_i, by comparing z_i with the other logits $p_i = softmax(\frac{z_i}{T})$, where T is a temperature parameter that is normally set to 1. A higher value for T means a softer probability distribution over classes.

In our work, the class probability of the ith region proposal predicted by the teacher model as $\{p_{i,l}^t, l \in L\}$, where l is the lth class probability and L is the number of categories. $p_{i,l}^t$ is treated as the soft targets to be transferred from the teacher to the student by optimizing the following cross entropy loss function as our classification distillation loss L_{cls_dis}:

$$L_{cls\_dis} = -\frac{1}{N} \sum_{i \in N} \sum_{l \in L} p_{i,l}^t log(p_{i,l}^s) \qquad (2)$$

where $p_{i,l}^s$ is the class probabilities predicted from student model, and N is the number of region proposals. The

soft targets contain information about the relationship between different classes as discovered by the teacher model. By learning from the soft targets, the student model inherits such hidden information.

For the bounding box regression branch which adjusts the location and size of the proposals, we define bounding box distillation loss using L2 optimization as shown below:

$$L_{bbox\_dis} = \frac{1}{2N}\sum_{i\in N}\|R_i^t - R_i^s\|_2^2 \quad (3)$$

where R_t and R_s means the outputs of regression layer from the teacher and student respectively. L_{bbox_dis} could encourage the student to get closer to the teacher's performance in terms of box regression.

Distillation for mask. The student network is trained with the help of a teacher in the mask prediction head with a channel-wise distillation. The prediction result of the mask head is a multi-channel mask. In each channel, the mask means the segmentation result of a certain category or simply a background or a foreground. Compared with semantic segmentation which can be treated as a pixel-wise classification task using *softmax* to predict the class for each pixel, instance segmentation decouples classification and segmentation heads to predict class-specific masks. Inspired by [11] which suggests that the competition among classes is not good for mask prediction, we designed the mask distillation loss L_{mask_dis} in channel-wise as shown below:

$$L_{mask\_dis} = -\frac{1}{2N\times C}\sum_{i\in N}\sum_{c\in C}\|M_{i,c}^t - M_{i,c}^s\|_2^2 \quad (4)$$

where $M_{i,c}^t$, $M_{i,c}^s$ are 14×14 mask outputs of the student and teacher network at channel c in the region proposal i in our setting. C denotes the channel number which equals to the number of categories. It is defined in the way such that the class-specific mask output of the student segmentation head layer is similar with that of the teacher network, regardless of the class prediction. Thus, our semantic distillation loss L_{sm_dis} is:

$$L_{sm\_dis} = L_{cls\_dis} + L_{bbox\_dis} + L_{mask\_dis} \quad (5)$$

which is the ensemble of each head distillation loss.

3.3. Hierarchical distillation (H-Dis)

Our overall distillation training loss is a hierarchical distillation loss which consists of representation and semantic distillation losses that can be written as L_{his_dis}:

$$L_{h\_dis} = L_{gt} + \lambda L_{sm\_dis} + \sigma L_{rp\_dis} \quad (6)$$

where the hyper-parameters λ and σ denote the balance parameters among different losses, which are fixed as 1 in our experiments.

4. Experiments

In this section, we perform evaluations on different backbones and different datasets to prove the effectiveness and generalization ability of our approach. In detail, we use FCIS[18] instance segmentation framework with ResNet[12] as our backbone. Results are reported on Pascal VOC 2012 [8] and Cityscapes [4]. Accuracy is evaluated by mean average precision at mask-level IoU (intersection-over-union) thresholds at 0.5 and 0.7. Speed is counted for all the processing of one image input on a single Nvidia 1080Ti GPU and we implement our approach on Mxnet [3].

4.1. Implementation details

We implement hierarchical distillation following a two-stage training strategy proposed in [17]. The first stage is to train a RPN network w/o mimicking the feature maps in the RoIs. The second stage is to train the head networks w/o distillation. Since achieving the highest accuracy is not our target, our implementation is not optimized with any accuracy enhancement tricks. The RPN anchors span 3 scales and 1 aspect ratio in all the experiments.

Training: We mostly follow the training setting of FCIS [18]. We train on a single GPU so the batch size is 1. For experiments on Pascal VOC, 240k iterations are performed where the learning rates are 10^{-3} and 10^{-4} in the first 160k and the last 80k iterations respectively. The iteration number is 144k for experiments on Cityscapes. In each batch, 128 proposals will back-propagate their gradients. All the teacher models and student models are performed on the same training set and test/val set.

Inference: At test time, forward propagation is performed on 300 proposals for one image in the first iteration, and then another 300 proposals are generated after the box regression branch. Our approach won't add any computation at test time, so that the speed of the student model would be the same with the baseline model.

4.2. Ablation study on Pascal VOC

Ablation experiments are performed on Pascal VOC. Following the protocol in [18, 10, 6], model training is performed on the train set, and evaluation is performed on the validation set, including objects of 20 categories. The training images are resized to have a shorter side of 600 pixels.

The ablation study results of our work and other conventional instance segmentation works are shown in Table 1. Upon these popular works in instance segmentation, we choose FCIS [18] as our teacher and student instance segmentation framework, because it's competitive in both speed and accuracy and has a good balance between them. We can observe that the FCIS framework with ResNet-18 backbone as our student model could achieve a very high speed at 16 fps with 10.2% accuracy drop compared with

ResNet-101 as our teacher model. Our proposed approach could enhance the accuracy of the tiny student model to 58.1%, which is close to the result of PFN [19], while it can retain a 16 fps speed compared to the 1 fps of PFN.

Model	$AP_{0.5}$(%)	speed(fps)
SDS[10]	49.7	<1
CFM[5]	60.7	<1
PFN[19]	58.7	1
MNC[6]	63.5	1
Mask R-CNN[11]	69.0	3
Teacher(ResNet-101)	65.7	6
Student(ResNet-18)	55.5	16
Student w/ H-Dis	**58.1**	**16**

Table 1. Comparison results between our proposed method and other popular instance segmentation works on Pascal VOC 2012 dataset. Teacher(ResNet-101) is the well-trained teacher network. Student(ResNet-18) is our student network trained without distillation as our baseline. Student w/ H-Dis is our proposed student network trained with our hierarchical distillation loss.

GT	Cls	Box	Mask	Mid	$AP_{0.5}$	$AP_{0.7}$
✓					55.5	36.1
✓	✓				55.8	36.1
✓	✓	✓			56.2	36.3
✓	✓	✓	✓		57.2	39.2
✓	✓	✓	✓	✓	58.1	42.3

Table 2. Effectiveness about different settings of components in our hierarchical distillation on Pascal VOC 2012 dataset. Mid denotes the representation distillation from middle layers.

The Table 2 shows different strategies for distillation to highlight the effectiveness of different losses. The loss of mask head shows the highest improvement (+1% for $AP_{0.5}$) to the result. The reason is that we utilize the multi-class mask predictions instead of class-agnostic mask predictions for distillation, which could include the full class-specific information obtained by the teacher. The improvement by distillation of classification and bounding box regression heads is weak because instance segmentation task is complicated and has multiple impact factors, merely distilling the logits can not transfer strong knowledge from the teacher. And unlike distillation for discrete categories, the bounding box regression outputs could provide very wrong guidance toward the student model. However, by combining these losses together, the teacher knowledge becomes influential enough to contribute to student network training and the accuracy enhancement becomes more promising (totally +2.6% for $AP_{0.5}$).

Model	student	w/ H-Dis	speed(fps)
ResNet-18	55.5	58.1(+2.6)	16
ResNet-18-4	39.0	42.5(+3.5)	19

Table 3. Comparison results of student models with different backbones on Pascal VOC 2012 dataset. The teacher model is ResNet-101 when H-Dis is used.

Model	AP[test]	AP[val]
Teacher(ResNet-101)	26.5	31.5
Student(ResNet-18)	16.5	18.6
Student w/ H-Dis	**18.1**	**20.5**

Table 4. Comparison results on Cityscapes dataset.

Besides ResNet-18, we also perform the experiment on a compressed ResNet-18 named ResNet-18-4, of which channel numbers of every layer are reduced to 1/4 compared to the original one. The Table 3 shows the accuracy enhancement and speed result of ResNet-18-4, which obtains 3.5% improvement compared with the baseline. This result suggests that a lighter and weaker student model could achieve more accuracy enhancement by learning from a well-trained teacher model. We could also conclude that more improvement can be gained by distilling from a better teacher model, since the larger gap they have, the more informative and effective knowledge the teacher transfers.

To show the effectiveness of our proposed H-Dis method, sample instance segmentation results from PSACAL VOC dataset are shown in Figure 2. The top 3 columns of images show that distilled model can segment the instances that are missed in the baseline results. The last 3 columns suggest that our method can reduce some false positive results.

4.3. Results on Cityscapes

We further report results on the Cityscapes dataset, using 2975 finely annotated images for training, 500 validation images and 1525 test images for evaluation. The images have a high resolution of 1024×2048 and we rescaled the image to have a shorter side of 512 pixels in both training and testing. The dataset involves 8 caregories for the instance segmentation task, within dominating number of samples for the *person* and *car* categories.

The evaluation results on Cityscapes dataset are shown in Table 4. As expected, the performance of the student model increases 1.6% and 1.9% AP respectively in test set and val set, suggesting the effectiveness of our proposed method. The improvement is not as obvious as the results on Pascal VOC dataset mainly because the teacher model is not strong enough to provide effective knowledge to the

Figure 2. Sample results on PSACAL VOC 2012 from **Baseline: Student(ResNet-18)** and **Ours: Student w/ H-Dis**. The corresponding instances between ground truth and results are printed with the same mask color. The mask that printed with other color denotes to be false positive results.

student. Combined with the results in Table 3, we can verify that when the teacher is not well-trained enough to teach a student, the enhancement degree will be lower or even none.

5. Conclusion

In this paper, we investigate knowledge distillation for training tiny instance segmentation networks through a teacher-student learning framework with hierarchical distillation loss. Knowledge distillation is used to transfer the knowledge through the head and middle layers, distilling the structure information from the cumbersome teacher network to the compact student network. In addition to classification and box regression distillation in our semantic distillation, we introduced a channel-wise distillation for the mask branch to achieve instance-level mask knowledge transfer. The experiments on Pascal VOC and Cityscapes datasets show that our approach outperforms the baseline model in accuracy while achieving a promising speed for real applications. Our experiment on backbones with different size verifies that our approach is increasingly effective when the gap between the teacher and student is larger.

References

[1] Jose M. Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In *NIPS*, 2016. 1

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Krishna Chandraker. Learning efficient object detection models with knowledge distillation. In *NIPS*, 2017. 2

[3] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *CoRR*, abs/1512.01274, 2015. 4

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 4

[5] Jifeng Dai, Kaiming He, and Jian Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015. 2, 5

[6] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 2, 4, 5

[7] Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando de Freitas. Predicting parameters in deep learning. In *NIPS*, 2013. 1

[8] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2009. 4

[9] Ross B. Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[10] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 2, 4, 5

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 4, 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2015. 4

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 1, 2

[14] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *CoRR*, abs/1511.06530, 2016. 1

[15] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018. 1

[16] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *CoRR*, abs/1608.08710, 2017. 1

[17] Quanquan Li, Shengying Jin, and Junjie Yan. Mimicking very efficient network for object detection. In *CVPR*, 2017. 1, 2, 4

[18] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017. 2, 4

[19] Xiaodan Liang, Liang Lin, Yunchao Wei, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2978–2991, 2017. 5

[20] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *CVPR*, 2019. 1, 2

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[22] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2015. 1, 2

[23] Yi Wei, Xinyu Pan, Hongwei Qin, Wanli Ouyang, and Junjie Yan. Quantization mimic: Towards very tiny cnn for object detection. In *ECCV*, 2018. 3

[24] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *NIPS*, 2016. 1

[25] Jiafeng Xie, Bing Shuai, Jianfang Hu, Jingyang Lin, and Wei-Shi Zheng. Improving fast segmentation with teacher-student learning. In *BMVC*, 2018. 1, 2

[26] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, 2017. 1

[27] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. Accelerating very deep convolutional networks for classification and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1943–1955, 2016. 1

[28] Xiangyu Zhang, Jianhua Zou, Xiang Ming, Kaiming He, and Jian Sun. Efficient and accurate approximations of nonlinear convolutional networks. In *CVPR*, 2015. 1