On Boosting Semantic Street Scene Segmentation with Weak Supervision

Panagiotis Meletis and Gijs Dubbelman

***Abstract*— Training convolutional networks for semantic segmentation requires per-pixel ground truth labels, which are very time consuming and hence costly to obtain. Therefore, in this work, we research and develop a hierarchical deep network architecture and the corresponding loss for semantic segmentation that can be trained from weak supervision, such as bounding boxes or image level labels, as well as from strong per-pixel supervision. We demonstrate that the hierarchical structure and the simultaneous training on strong (per-pixel) and weak (bounding boxes) labels, even from separate datasets, consistently increases the performance against per-pixel only training. Moreover, we explore the more challenging case of adding weak image-level labels. We collect street scene images and weak labels from the immense Open Images dataset to generate the OpenScapes dataset, and we use this novel dataset to increase segmentation performance on two established per-pixel labeled datasets, Cityscapes and Vistas. We report performance gains up to +13.2% mIoU on crucial street scene classes, and inference speed of 20 fps on a Titan V GPU for Cityscapes at 512 x 1024 resolution. Our network and OpenScapes dataset are shared with the research community.**

I. Introduction

Semantic segmentation of street scene images is a fundamental building block for automated driving [1]. It is the first step of scene understanding and provides the necessary information towards higher level reasoning and planning [2]. Formulation of the problem as per-pixel (dense) classification and modeling it with Fully Convolutional Networks [3] has become the de facto solution for semantic segmentation of images. However, its success is based on the availability of huge amounts of tediously, per-pixel labeled datasets [4], [5], [6], and existing solutions do not leverage weakly labeled data that are provided in larger and more diverse datasets [7].

Therefore, in this work, we explore a method for per-pixel training of Fully Convolutional Networks on multiple datasets simultaneously, containing images with strong (per-pixel) or weak (bounding boxes and image-level) labels. The ability to train from weakly labeled data is an important research topic in the field of computer vision [8], [9], [10], [11], which, when solved, can be of benefit to many application domains.

The challenge when using weak supervision for per-pixel semantic segmentation lies on the different and incompatible *annotation types* [12]. Our method fully solves, in a consistent and uniform manner that challenge, while training on heterogeneous datasets. It consists of a hierarchy of classifiers and a hierarchical loss function, which can handle any of the aforementioned types of supervision. This is achieved by transforming the incompatible, for semantic segmentation, weak labels into per-pixel weak labels, a common practice also appearing in other related problems [13].

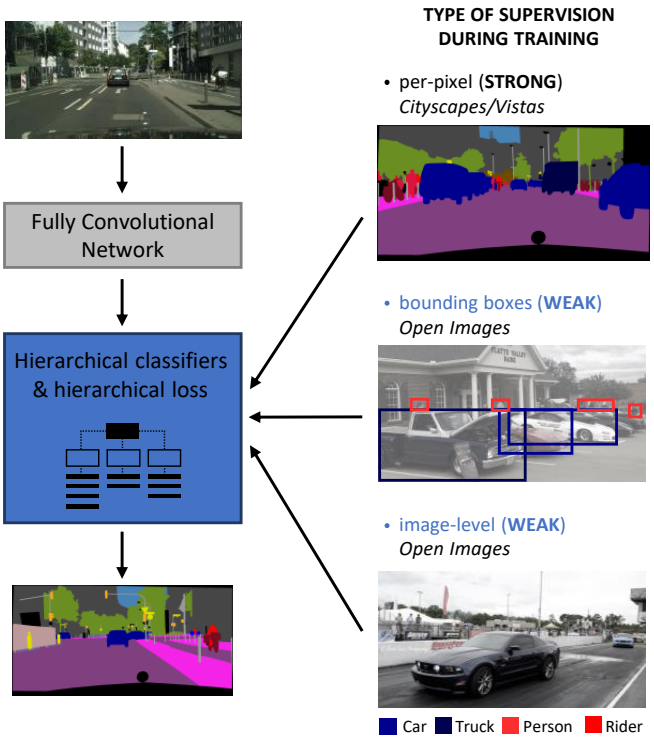

Fig. 1. Our contributions in blue color. Using diverse types of weak supervision from the Open Images dataset we achieve increase on performance over datasets with strong supervision using a fully convolutional network with a hierarchy of classifiers and the corresponding hierarchical loss.

In order to prove how weak labels can boost semantic segmentation performance of strongly (per-pixel) labeled datasets, *e.g.* Cityscapes [4], we collect a weakly labeled dataset by mining street scene images from the very large scale Open Images dataset [7] and we name it *OpenScapes*. This new dataset contains 100,000 images with 2,242,203 bounding box labels and 100,000 images with 1,199,582 image-level labels, and spans 14 of the most important street scene classes. However, even after the automated selection procedure, the *domain gap* between *OpenScapes* and the per-pixel datasets, for which we want to prove the performance increase, remains large, as can be seen in Fig. 3, making the

Panagiotis Meletis (p.c.meletis@tue.nl) and Gijs Dubbelman (g.dubbelman@tue.nl) are with the Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands.
This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 688099.

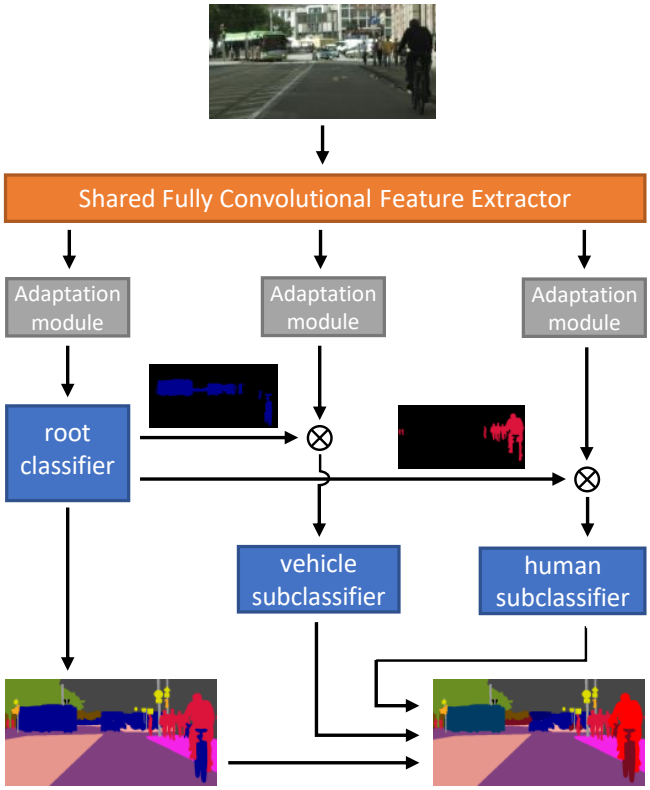

Fig. 2. Network architecture. The root classifier passes its decisions to the two subclassifiers, which classify only pixels that are assigned to them by the root classifier.

weak supervision even "weaker".

We evaluate our system on two established per-pixel labeled datasets and we show that performance increase is proportional to the amount of extra weak labels used. We achieve that without using any external components as in [11], [8], but only the hierarchical structure of the classifiers and the proposed hierarchical loss.

To summarize, the contributions of this work are:

- A methodology for training semantic segmentation networks on datasets with diverse supervision, including per-pixel, bounding box, and image-level labels.
- *OpenScapes* dataset: a large, weakly labeled dataset with $200,000$ images and 14 semantic classes for street scenes recognition.

Our system and the *OpenScapes* dataset are made available to the research community [14].

II. Method

In this Section, we describe the proposed training and inference methodology. We generalize previous work [12] by enabling weak supervision from both bounding box labeled and image-level labeled images without using any external components.

Our method facilitates training of any fully convolutional network for per-pixel semantic segmentation and only requires a specific structure of classifiers and a specialized

Fig. 3. Example images from per-pixel labeled Cityscapes dataset and the weakly labeled *OpenScapes* dataset that demonstrate the big *domain gap*.

loss to train them. To achieve that, weak labels (bounding boxes and image labels) have to be converted to pseudo per-pixel ground truth. This is described in Sec. II-B. The network architecture and the corresponding hierarchical loss are presented in Sec. II-A and Sec. II-C respectively. In addition, we address the shortcomings of pseudo ground-truth generation [12] for any type of weak labels.

A. Convolutional Network Architecture

The network architecture follows the design proposed in [12] and is depicted in Fig. 2. Specifically, we opt for a two-level hierarchical convolutional network, which consists of a fully convolutional shared feature extractor and a set of, hierarchically arranged, classifiers. The root classifier is trained only with strong supervision (per-pixel labeled semantic classes). The subclassifiers are trained, using the hierarchical loss of Sec. II-C with per-pixel supervision. For that purpose, the weak labels are converted into per-pixel pseudo ground truth as described in Sec. II-B.

The benefits of the hierarchical structure [12] are twofold: 1) it solves the problem of simultaneously training with different types of supervision, by placing classes with weak labels in the subclassifiers, and 2) it solves the semantic class incompatibilities between datasets, due to the unavailability of specific semantic classes in all datasets.

The hierarchy of classifiers is constructed according to the availability of strong and weak labels for each class. The root classifier (left in Fig. 2) contains high-level classes with per-pixel labels. Each one of the subclassifiers corresponds to one high-level class of the root classifier, and contains subclasses with per-pixel and/or weak supervision.

The shared feature representation (see Fig. 2) is passed through two shallow, per-classifier *adaptation networks*, which adapt the common representation, its depth, and receptive field to meet the requirements of each classifier as described in [12]. In this work, we use a single ResNet bottleneck layer [15] as in [12].

B. Generation of pseudo per-pixel ground truth from weak labels

The goal is to train the network with per-pixel labels, thus we need to generate per-pixel ground truth from bounding boxes and image-level labels. The 2D ground truth generation procedure in [12] is ambiguous for classes whose bounding box boundaries do not match tightly the object boundaries. Thus that method is valid only for square-shaped,

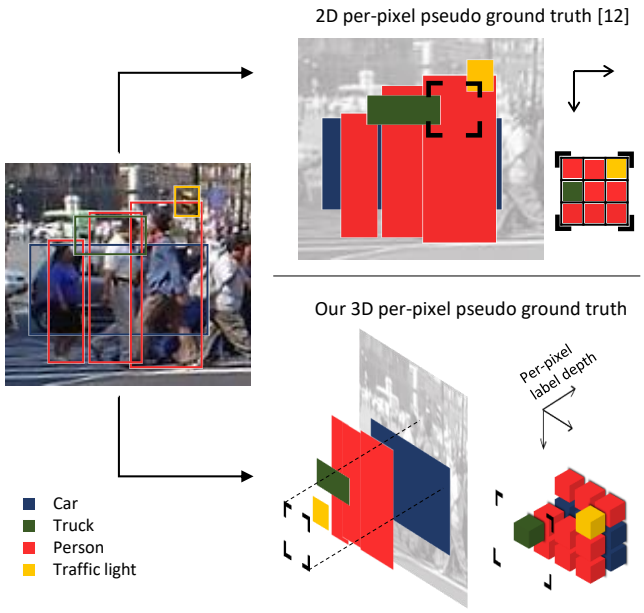

Fig. 4. 2D vs 3D per-pixel pseudo ground truth (GT) generation. Left: image with a selected subset of bounding boxes colored by class. Right top: 2D GT generation used in [12]. Right bottom: proposed 3D GT generation. In 2D GT, overlapping bounding boxes produce ambiguity in generated per-pixel labels (*e.g.* the car label is hidden behind pedestrian labels), which is solved by adding a 3rd dimension in GT generation. The pseudo ground truth has the form of a dense categorical instead of a sparse (one-hot) distribution. The same principle is used for generating GT from image level labels by considering for each label the boundaries of the bounding box to extend to the whole image.

compact objects, like traffic signs, and cannot be applied to image-level labels.

According to [12] the 2D pseudo ground truth for each image is generated pixel-wise, by assigning a single label to each pixel from the set of bounding boxes that this pixel belongs to, as depicted at the top of Fig. 4. This procedure effectively generates a so-called sparse or one-hot categorical probability distribution, since each pixel belongs to a specific class with probability 1. Contrary, in this work, we model the per-pixel labels as a dense or multi-hot categorical probability distribution, and thus the ground truth for each images becomes 3D (see Fig. 4). This model assigns to each pixel a probability for every class, and the sum of probabilities for all classes must be 1. In order to convert bounding boxes and image labels to per-pixel labels, we use a voting scheme, according to which each label increases each pixel's counter vector by 1. After collecting all votes we normalize across all classes, in order for the labels to represent a valid probability distribution.

C. Hierarchical loss

We construct the hierarchical loss similar to [12], namely the loss is accumulated unconditionally for per-pixel labeled datasets and conditionally for per bounding box or per image-level labeled datasets. The total loss terms are pixel-wise categorical cross entropy losses and are summarized in Table I. The five loss terms of Table I are added, using

TABLE I

LOSS COMPONENTS PER CLASSIFIER AND PER DATASET. ALL LOSSES ARE PER-PIXEL CATEGORICAL CROSS ENTROPY (CCE) LOSSES BETWEEN THE DENSE OR SPARSE CATEGORICAL LABELS AND THE SOFTMAX PROBABILITIES OF THE ASSOCIATED CLASSIFIER.

classifier	Per-pixel labeled data (Cityscapes or Vistas)	Weakly labeled data (OpenScapes)
root	sparse CCE	-
vehicle subcl.	dense CCE	conditional dense CCE
human subcl.	dense CCE	conditional dense CCE

TABLE II

OPENSCAPES DATASET OVERVIEW AND COMPARISON WITH PER-PIXEL LABELED DATASETS. TRAINING SPLITS ARE SHOWN.

	Cityscapes	Vistas	OpenScapes
# of images	2975	18000	200,000
# of classes	27	65	14
# of pixel labels	$1.6 \cdot 10^9$	$156.2 \cdot 10^9$	-
# of bound. boxes	-	-	2,242,203
# of image labels	-	-	1,199,582

coefficients of 0.1 for the subclassifier's losses and, together with the regularization loss, make up the total loss.

We assign a set of n classes named $\{1, ..., n\}$ for each classifier. The per-pixel labels are given in the form of a vector $\mathcal{Y} = [y_1, ..., y_n]$ with elements corresponding to per-class probabilities from a categorical distribution, *i.e.* $\sum_{y\ in\ \mathcal{Y}} y = 1$. The general form of the per-pixel categorical cross-entropy loss for a softmax classifier with n classes and softmax output $\boldsymbol{\sigma} = [\sigma_{y_1}, ..., \sigma_{y_n}]$ for all pixels $\mathcal{P}$ is:

$$\mathcal{L} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{y \in \mathcal{Y}} y \log \sigma_y \tag{1}$$

For the root classifier, we use only the sparse categorical per-pixel labels, since this classifier receives supervision from the per-pixel labeled dataset. In this case, $y_j = 1$ for the class j that the pixel is labeled with, and $y_{i \neq j} = 0$ for all other classes. Thus, Eq. 1 is reduced to sparse CCE loss:

$$\mathcal{L} = -\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \log \sigma_{y_j} \tag{2}$$

For the subclassifiers, we use dense categorical per-pixel labels for both the per-pixel and the weakly labeled images (see Sec. II-B). We convert the per-pixel labeled dataset's sparse labels to dense categorical labels by assigning a probability of $y_j = 1$ to the ground truth class j, and probability $y_{i \neq j} = 1$ to all other classes. For per-pixel labeled images we use Eq. 2. For weakly labeled images we use the loss of Eq. 1, which is accumulated from pixels $\mathcal{P}$ that satisfy two conditions: 1) the per-pixel pseudo ground truth has non-zero probability for that pixel, and 2) the root classifier decision agrees with the per-pixel pseudo ground truth for that pixel, *i.e.* it is a class that has non-zero probability in the per-pixel pseudo ground truth.

III. OpenScapes Dataset and Implementation

In this Section, we describe the collection process of *OpenScapes* and we compare it with the per-pixel annotated Cityscapes and Vistas datasets. Moreover, we discuss all the implementation details for our experiments.

A. OpenScapes Street Scenes Dataset

We collect images of street scenes from the recently open-sourced, very large scale Open Images dataset [7] and create a subset that we call *OpenScapes*. Open Images dataset contains over 9,000,000 images, 14,600,000 bounding boxes for 600 object classes, and more than 27,900,000 human-verified image-level labels for 19,794 classes. We collected 200,000 images, containing 2,242,203 bounding box labels and 1,199,582 image-level labels from 14 classes, with as much as possible street scene related content.

The fully automated collection procedure is described in Sec. III-A.1. However, even after the careful selection, the *domain gap* [16], [17] between the per-pixel datasets (Cityscapes, Vistas) and *OpenScapes* is large. This can be seen by the image examples in Figures 1 and 3, and is discussed together with a comparison with the employed per-pixel labeled datasets in Sec. III-A.2.

1) ***Mining procedure****:* First, we rank in descending order images from Open Images by the number of bounding boxes and image-level labels they contain for the 14 selected street scene classes. Then we select the top 100,000 images for the bounding box labeled subset and then 100,000 images for the image-level labeled subset and we make sure that there is no image overlap between the two subsets. For the ranking we used a voting system, according to which classes in the weak labels of an image vote for an image to be a street scene image or not. The more probable classes, like traffic light and license plate, can cast more votes than classes that may appear in other contexts (*e.g.* car, person).

2) ***Comparison with per-pixel labeled datasets****:* In Tab. II we compare *OpenScapes* with two established per-pixel labeled datasets that we experiment on also in this paper. In Fig. 3 we present some images from Cityscapes and *OpenScapes*. As can be seen Cityscapes image domain is very consistent with images taken from a specific point of view and in one country, contrary to *OpenScapes*, which contains web-like images and does not correspond to a consistent domain.

B. Implementation details

The network is depicted in Fig. 2. The feature extractor consists of the ResNet-50 layers (without the classifier) from [15], followed by an 1x1 convolutional layer, to decrease feature dimensions to 256, and a Pyramid Pooling Module [18]. The stride of the feature representation on the input is reduced from 32 to 8, using dilated convolutions. Each branch has an extra bottleneck module [15], a bilinear upsampling layer to recover original resolution, and a softmax classifier.

We use Tensorflow [19] and 4 Titan V 12 GB GPUs for training. We implemented synchronous, cross-GPU batch normalization, and for all experiments we use batch size of up to 4 images per-GPU depending on the experiment, containing 1 image from the per-pixel labeled dataset (Cityscapes or Vistas), 2 images from the bounding box labeled dataset (*OpenScapes* subset), and 1 image from the image-level labeled dataset (*OpenScapes* subset).

For experiments involving Cityscapes we use images dimensions of 512x1024 and for Vistas 621x855. Since, *OpenScapes* images have multiple aspect ratios, we upscale each image to fit tightly the aspect ratio of the per-pixel labeled dataset and then we crop a random patch of same dimensions as the per-pixel labeled image. Networks with batch size of 3 per-GPU are trained for 26 epochs with initial learning rate 0.02 and of 4 per-GPU for 31 epochs with initial learning rate 0.03. All networks are trained with Stochastic Gradient Descent and momentum of 0.9, L2 weight regularization with decay of 0.00017, the learning rate is halved three times, and batch normalization moving averages decay set to 0.9. We use the same hyperparameter values for the $\lambda^j = 0.1$ coefficients of the loss as in [12].

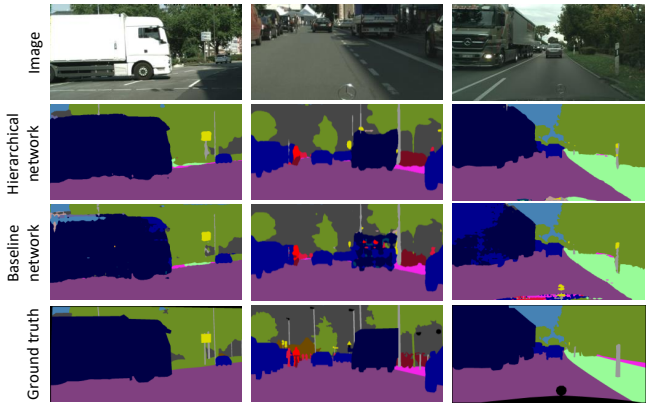

Fig. 5. Comparison for Cityscapes validation split images for the hierarchical network trained on *OpenScapes* and Cityscapes against the baseline network trained on Cityscapes only. The three classes with biggest improvement in mIoU are Truck(+13.2%), Rider(+3.5%), and Person(+2.1%).

TABLE III
OVERALL PERFORMANCE IMPROVEMENTS USING WEAK SUPERVISION FROM *OpenScapes* DATASET IN ADDITION TO STRONG SUPERVISION FROM CITYSCAPES OR VISTAS, OVER THE BASELINE NETWORK TRAINED WITH ONLY PER-PIXEL LABELS.

Origin of ground truth			results on val splits			
Cityscapes	OpenScapes		Cityscapes		Vistas	
per-pixel	bound. boxes	image-level	mAcc	mIoU	mAcc	mIoU
✓			77.8	68.9	53.0	43.6
✓	✓		79.2	70.2	52.1	43.6
✓	✓	✓	79.3	70.3	52.0	43.0

IV. Experiments

We evaluate performance using two established multi-class metrics for semantic segmentation [4], namely mean pixel Accuracy (mAcc) and mean Intersection over Union (mIoU). Metrics for all experiments are evaluated on the

TABLE IV

CITYSCAPES PER CLASS **MIOU (%)** IMPROVEMENTS, FOR THE CLASSES, WHICH BELONG TO SUBCLASSIFIERS THAT RECEIVE EXTRA SUPERVISION FROM THE WEAKLY LABELED *OpenScapes* DATASET (100K SUBSETS). RESULTS ARE GROUPED PER SUBCLASSIFIER.

Origin of ground truth			Vehicle subclassifier							Human subclassifier		
Cityscapes per-pixel	OpenScapes bound. boxes	OpenScapes image-level	Bicycle	Bus	Car	Motorc.	Train	Truck	Overall	Person	Rider	Overall
✓			67.0	79.7	91.9	**52.2**	**69.3**	62.3	70.4	70.2	47.9	59.0
✓	✓		67.8	**81.8**	**92.5**	50.3	**69.3**	71.4	**72.2**	71.9	50.7	61.3
✓	✓	✓	**67.9**	79.1	**92.5**	48.7	**69.3**	**75.5**	**72.2**	**72.3**	**51.4**	**61.9**

validation splits (Cityscapes: 500, Vistas: 2000) of the per-pixel datasets, and are averaged on the last three epochs. In Sec. IV-A and IV-B we present overall results and per class results for the classes that receive extra weak supervision. In Sec. IV-C we investigate the effect of the number of examples used from the weakly labeled dataset. Example results from all datasets are shown in Figures 5 and 6.

A. Overall results

In Table III the overall results for Cityscapes [4] and Vistas [5] are shown. All networks are trained with strong (per-pixel) supervision, from Cityscapes or Vistas, and a combination of weak (per bounding box or image-level or both) supervision from *OpenScapes*. We used the two subsets of *OpenImages* with 100k images each (Sec. III-A) and their generated pseudo per-pixel labels, as described in Sec. II-B, mixed in the batch with Cityscapes or Vistas images (see Sec. III-B for implementation details).

For Cityscapes, we observe that mAcc and mIoU increase steadily by increasing the amount of weakly labeled data included during training. For Vistas, however, training together with the *OpenScapes* subsets slightly harms the performance. This is possibly due to the diversity of images of Vistas and the large *domain gap* with *OpenScapes*. Overall, we denote that by adding extra supervision for specific classes, mean performance over all classes is not harmed dramatically, and in most cases also boosted.

B. Improvements on classes with weak supervision

In this Section, we investigate the performance on classes that receive extra weak supervision apart from strong per-pixel supervision. As can be seen in Tables IV and V, overall mIoU of classes belonging to vehicle and human subclassifiers improves in both datasets when adding the *OpenScapes* bounding box labeled subset. Although, in the Cityscapes case, adding the *OpenScapes* image-level labeled subset increases the performance, in the Vistas case it reduces it. We hypothesize that this is due to the *domain gap* between the datasets (see Sec. III-A.2, V). We would also like to mention the big increase for specific classes, *e.g.* +13.2% for Cityscapes "Truck" class, +11.3% for Vistas "Caravan" class, and +10.8% for Vistas "On rails" class.

C. Effect of weakly labeled dataset size

In this experiment we train the hierarchical architecture on Cityscapes, together with different portions of the *OpenScapes* bounding boxes labeled subset, with all other hyperparameters fixed, to investigate the effect of the size of the weakly labeled dataset. From Table VI, row 2, it becomes clear that without using enough weakly labeled images the performance may even drop. However, when enough weak supervision is provided, row 3 and 4, the performance is enhanced adequately.

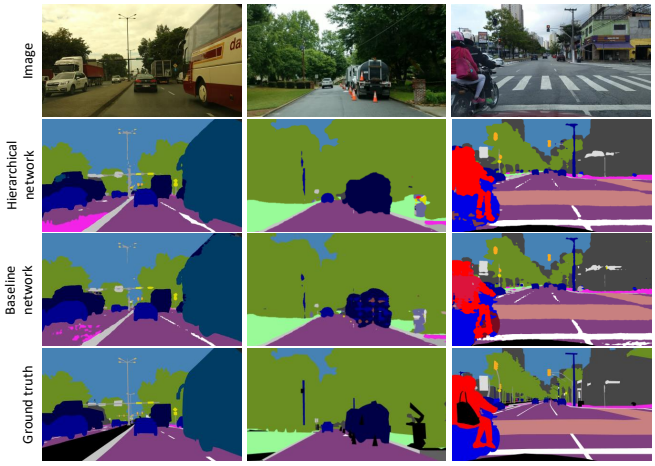

Fig. 6. Comparison for Vistas validation split images for the hierarchical network trained on *OpenScapes* bounding boxes subset and Vistas against the baseline network trained on Vistas only. The three classes with biggest improvement in mIoU are Caravan(+11.3%), On rails(+10.8%), and Motorcyclist(+9.6%).

V. DISCUSSION AND FUTURE WORK

The performance of our method heavily depends on two factors: 1) the amount of weak labels and their semantic extent of class connotation, and 2) the *domain gap* [16], [17], [20] between strongly and weakly labeled datasets.

In this work we hypothesized that the images for datasets that are trained simultaneously come from similar domains, and thus features from a common feature extractor can be classified by the same classifier. In reality, this assumption rarely holds, but we leave investigation of this matter and how to solve it during inference to future research. Methods that perform domain agnostic inference, like [20], can hold solutions for this problem.

Another important matter is the *connotation extent* (the extent of the class name connotation for labeling visually similar objects) of a semantic class. Although in this work,

TABLE V

VISTAS PER CLASS **MIOU** (%) IMPROVEMENTS, FOR THE CLASSES, WHICH BELONG TO SUBCLASSIFIERS THAT RECEIVE EXTRA SUPERVISION FROM THE WEAKLY LABELED *OpenScapes* DATASET (100K SUBSETS). RESULTS ARE GROUPED PER SUBCLASSIFIER.

Origin of ground truth			Vehicle subclassifier											Human subclassifier				
Cityscapes per-pixel	OpenScapes bound. boxes	OpenScapes image-level	Bicycle	Boat	Bus	Car	Caravan	Motorcycle	On rails	Other veh.	Trailer	Truck	Wheeled	Overall	Person	Cyclist	Motorcyc.	Other rider	Overall
✓			55.0	**26.7**	**75.0**	**88.8**	0.3	**54.2**	38.4	16.9	0.3	65.0	7.4	38.9	**65.5**	**51.4**	43.1	0.0	40.0
✓	✓		**56.1**	21.2	73.8	88.6	**11.6**	53.9	**49.2**	**18.4**	**0.9**	**66.9**	**10.7**	**41.0**	64.7	47.1	**52.7**	**0.4**	**41.2**
✓	✓	✓	54.5	21.2	74.0	88.4	11.4	52.8	49.0	18.1	0.8	66.0	10.6	40.6	64.6	47.1	49.9	0.3	40.5

TABLE VI

PERFORMANCE (%) ON CITYSCAPES WITH DIFFERENT AMOUNT OF BOUNDING BOXES USED TO GENERATE PSEUDO GROUND TRUTH LABELS FOR THE WEAKLY LABELED DATASET.

per-pixel + *#images* with bbox GT	mAcc	mIoU
0 images (0 bboxes)	77.8	68.9
$1k$ images ($17.3k$ bboxes)	77.4	68.4
$10k$ images ($140.4k$ bboxes)	78.2	69.2
$100k$ images ($1185.8k$ bboxes)	79.2	70.2

we assumed that classes described by the same high-level semantic concepts, like truck or bus, depict very similar objects across datasets, this is not true in general, and should be investigated in the future. This is visible, for example, in the performance drop for the motorcycle class in Table IV, for which the *connotation extent* for motorcycle objects diverges between Cityscapes and *OpenScapes* datasets.

VI. CONCLUSION

We presented a fully convolutional network coupled with a hierarchy of classifiers for simultaneous training on strongly and weakly labeled datasets for semantic segmentation. We collected street scene images from Open Images to generate a weakly labeled dataset called *OpenScapes*. Using *OpenScapes* we showed that the overall performance, as well as the performance for classes that receive extra weak supervision, is increased, provided that enough weak labels are available. Moreover, we examined the effect of the size of the weakly labeled dataset and showed that the performance increase is proportional to the size of the dataset. For our experiments we assumed that the *domain gap* between simultaneously trained datasets is minor, however in other cases it can be a limiting factor, especially when using image-level labels, and should receive attention in future research.

REFERENCES

[1] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art," *CoRR*, vol. abs/1704.05519, 2017.

[2] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 187–210, 2018. [Online]. Available: https://doi.org/10.1146/annurev-control-060117-105157

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[4] 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] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy*, 2017, pp. 22–29.

[6] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018, pp. 954–960.

[7] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *arXiv:1811.00982*, 2018.

[8] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1742–1750.

[9] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 695–711.

[10] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[11] L. Ye, Z. Liu, and Y. Wang, "Learning Semantic Segmentation with Diverse Supervision," *arXiv preprint arXiv:1802.00509*, Feb. 2018.

[12] P. Meletis and G. Dubbelman, "Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation," in *IEEE IV 2018*. IEEE, 6 2018, pp. 0–8.

[13] C. Lu, M. J. G. van de Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoderdecoder networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 445–452, April 2019.

[14] "Full implementation code for training, evaluation and inference, and the extra annotated datasets will be made publicly available at https://github.com/pmeletis/hierarchical-semantic-segmentation."

[15] 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.

[16] J. Zhang, W. Li, and P. Ogunbona, "Transfer learning for cross-dataset recognition: a survey," *arXiv preprint arXiv:1705.04396*, 2017.

[17] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arXiv preprint arXiv:1702.05374*, 2017.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017.

[19] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[20] R. Romijnders, P. Meletis, and G. Dubbelman, "A domain agnostic normalization layer for unsupervised adversarial domain adaptation," *CoRR*, vol. abs/1809.05298, 2018. [Online]. Available: http://arxiv.org/abs/1809.05298