SEMANTIC IMAGE SEGMENTATION GUIDED BY SCENE GEOMETRY

Sotirios Papadopoulos, Ioannis Mademlis, Ioannis Pitas

Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki, Greece

ABSTRACT

Semantic image segmentation is an important functionality in various applications, such as robotic vision for autonomous cars, drones, etc. Modern Convolutional Neural Networks (CNNs) process input RGB images and predict per-pixel semantic classes. Depth maps have been successfully utilized to increase accuracy over RGB-only input. They can be used as an additional input channel complementing the RGB image, or they may be estimated by an extra neural branch under a multitask training setting. Contrary to these approaches, in this paper we explore a novel regularizer that penalizes differences between semantic and self-supervised depth predictions on presumed object boundaries during CNN training. The proposed method does not resort to multitask training (which may require a more complex CNN backbone to avoid underfitting), does not rely on RGB-D or stereoscopic 3D training data and does not require known or estimated depth maps during inference. Quantitative evaluation on a public scene parsing video dataset for autonomous driving indicates enhanced semantic segmentation accuracy with zero inference runtime overhead.

***Index Terms*—** semantic segmentation, depth estimation, scene geometry, computer vision

1. INTRODUCTION

Semantic image segmentation is one of the most essential scene understanding tasks in modern computer vision, mainly due to its critical importance for autonomous systems, robots and vehicles [1, 2, 3]. It consists in classifying each input image pixel into one amongst a set of prespecified object classes. Convolutional Neural Networks (CNNs) have been the state-of-the-art in similar perception tasks for a long time now. Traditionally, single-view RGB footage has been considered as an adequate input modality for CNNs to successfully perform semantic segmentation. However this is not always the case, since in certain scenarios individual RGB images fail to provide sufficient class-distinctive hints.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871479 (AERIAL-CORE).

Scene/object geometry has been long known to provide insightful information on several computer vision tasks. Geometry can be described by various formats, such as depth maps, since it can provide cues about shape, texture and distance from the image plane. However, the acquisition of depth maps when constructing an application-specific dataset (e.g., by a depth camera, a LIDAR sensor, etc.) can be a cumbersome task. Typically, high-accuracy depth sensors are expensive and their outputs need heavy post-processing. To alleviate this, a lot of unsupervised/self-supervised depth estimation CNNs have been recently proposed [2]. Such methods learn to infer depth from monocular RGB images by depth supervision [4], by stereo parallax estimation [5], or, as of lately, via monocular video sequences under a neural Structure-from-Motion (SfM) paradigm [6].

The trivial way to exploit this for improving semantic segmentation would be an inference-time two-stage, multimodal approach: estimate a depth map per image/video frame, pair it with corresponding RGB data and feed to it to a neural segmentor pre-trained on RGB-D inputs. Thus, maximum information is extracted from the input image in a pre-processing step performed on-the-fly, facilitating the succeeding CNN in its semantic segmentation task. However, this comes at a significant runtime penalty during model deployment, demanding two different CNNs to be executed in series. An alternative approach would be to train a multitask CNN that concurrently performs both semantic segmentation and depth map estimation from RGB input, using two task-specific neural heads and a common backbone CNN for feature extraction, but this increases training difficulty and requires more complex, slower CNN architectures, able to handle both tasks simultaneously.

This paper presents a novel regularizer that utilizes neurally-estimated (i.e., without requiring any dedicated sensors) depth maps, only while training a conventional semantic segmentation CNN, using regular RGB video data for training (not RGB-D or stereoscopic 3D) and without resorting to a multitask setting (which could demand higher model complexity to avoid underfitting). Thus, the proposed regularizer exploits the depth map modality for increasing segmentation accuracy, without imposing *any* runtime overhead during model deployment/inference and without relying on special

input data types at *any* stage. It operates by penalizing differences between semantic and depth predictions on presumed object boundaries. Depth maps are estimated using a separate neural branch, pretrained on regular RGB videos in a self-supervised manner and totally independent from the main segmentation CNN. After training is complete, the latter one can be employed alone, for processing individual, previously unseen single-view RGB images, without relying on dedicated depth sensors or separate geometry estimation CNNs.

Quantitative evaluation of the presented method on a public, scene parsing video dataset for autonomous driving yielded favourable results, in comparison both to the baseline semantic segmentation CNN and to competing methods.

2. RELATED WORK

A vast amount of previous work [7, 8] treats depth as a given input modality to perform computer vision tasks. However these approaches require scene geometry to be known (e.g., by relying on RGB-D sensors), therefore they are not directly related to this paper.

In a number of more relevant cases, depth maps are estimated from RGB input along with semantic segmentation maps, so as to increase accuracy, under a multitask training setting and, typically, with supervised depth estimation (i.e., ground-truth RGB-D data are required during training) [9, 10]. However, certain algorithms falling under this category employ semantic segmentation for extracting improved depth maps, instead of the reverse, and rely on self-supervised depth estimation using stereoscopic 3D images, instead of RGB-D data (e.g., [11, 12, 13]). In [13], enhanced consistency between the two tasks is achieved by inserting a *Cross-Domain Discontinuity* loss term during training, based on the observation that depth discontinuities are likely to co-occur with semantic boundaries. This term detects discontinuities between semantic labels by computing the sign of the absolute value of the gradients in the semantic map. The underlying intuition is that there should be a gradient peak between neighboring pixels belonging to different classes. [12] is a different multitask architecture for semantic image segmentation that is also trained with a similar smoothness loss term. Unlike [13], where the Cross-Domain Discontinuity Term enforces smoothness on the depth values within each ground-truth segmentation mask (thus no error gradients propagate through the segmentation decoder), the regularizer proposed in [12] is computed based on the segmentation branch output. The multitask network shares both the encoder and the decoder, differentiating only in the prediction heads. The decoder is given a task identity signal to predict features for either task.

Certain similar methods employ a pretrained semantic segmentation network to improve depth estimation accuracy, instead of joint multitask training. In two-stage approach [11], depth maps are estimated from stereoscopic 3D image pairs and, subsequently, depth borders are optimized using prior-predicted semantic borders. Then, the method explicitly morphs the predicted depth maps so that depth edges coincide with semantic edges, while finally using this improved depth information as a supervision signal. [14] uses a pretrained segmentor's features to guide a self-supervised depth estimation network's decoder via pixel-adaptive convolutions. Depth estimation relies on video data and on a SfM training loss function.

Few papers report semantic segmentation performance gains by exploiting self-supervised depth estimation. [15] shows that the encoder can have a better weight initialization than simply pretraining on the ImageNet dataset for whole-image classification, by pretraining on automatically computed relative depth derived from self-supervised optical flow; thus the method employs a two-stage training process requiring video data. On the contrary, most other similar approaches rely on a multitask training setting. A number of these multitask methods need stereoscopic 3D training data, such as [16], which trains a multitask network for semantic segmentation, self-supervised depth estimation and image colorization to enhance semantic segmentation performance. On the other hand, [17] estimates depth in a self-supervised manner from regular videos and trains the CNN under a multitask setting with task-specific decoders, achieving a substantial performance increase. [18] also leverages multitask training and self-supervised monocular depth estimation from monocular videos to improve semantic segmentation performance, but it is designed for the special case of semi-supervised learning; thus, it is not directly related to this paper.

Focusing only on papers most similar to ours, the regularizers presented in both [13] and [12] are used mainly to guide depth estimation under a multitask training setting. [16] also employs a multitask architecture, to improve semantic segmentation by exploiting depth estimation. In comparison, the regularizer proposed in this paper is employed for optimizing fully supervised semantic segmentation, without resorting to a multitask architecture or training. Thus, from a different perspective and at a high level of abstraction, the proposed method can be broadly seen as the inverse of [11], which is designed for depth estimation.

3. GEOMETRY-GUIDED SEMANTIC SEGMENTATION

The proposed method does not require RGB-D or stereoscopic 3D training data (which may be difficult to acquire for specific applications), does not impose any runtime overhead during inference on the trained model (as the naive inference-time two-stage approach does) and requires no architectural modifications to the semantic segmentation CNN for facilitating multitask training without underfitting. It consists in: a) self-supervised pretraining of a separate depth map estimation CNN branch and, b) subsequently, training in a regular

manner any conventional semantic segmentation CNN with an additional regularizing loss term, i.e., the proposed *holistic consistency* loss. The latter one is computed at each training iteration using the outputs of the segmentor and of the pretrained depth estimation branch.

The underlying intuitive observation was that semantic objects tend to stand out in depth maps, leading to co-occurence of image gradients in the two tasks. The proposed loss term penalizes semantic map edges that are absent from the spatially corresponding region of the respective depth map, since the target is to enhance semantic segmentation accuracy using depth and not vice versa. As a result, during training, the segmentor is discouraged from outputting semantic shapes that do not conform to scene geometry. Thus, depth information is implicitly integrated while the CNN model is being optimized, but, subsequently, no depth inputs or depth estimation neural branches are required during inference. It is clear that the depth estimation branch can be totally omitted in model deployment, since it is only required during training for computing the proposed regularizer.

3.1. Notation

- $C \in \mathbb{N}$: the number of semantic classes.
- $N_1, N_2 \in \mathbb{N}$: the image spatial dimensions (in pixels).
- $\{A(i,j)\}_{\substack{1 \le i \le N_1, \\ 1 \le j \le N_2}}$: A matrix $\mathbf{A} \in \mathbb{R}^{N_1 \times N_2}$, composed of entries A_{ij}, $1 \le i \le N_1, 1 \le j \le N_2$.
- $\mathbf{S} \in \mathbb{R}^{N_1 \times N_2 \times C}$: the estimated segmentation map. It is a tensor, with each of its C 2D channels being class probability heat maps.
- $\mathbf{D} \in \mathbb{R}^{N_1 \times N_2}$: the estimated depth map. It is a matrix containing the normalized distance of each depicted 3D point from the image plane.
- $mean(\mathbf{A})$: the mean over all entries of matrix $\mathbf{A}$.
- $max(\mathbf{a})$: the maximum over all entries of vector $\mathbf{a}$.
- x, y: the spatial image axes.

3.2. Per-class consistency loss

Here we introduce a preliminary, more involved variant of the proposed holistic consistency loss, in order to facilitate understanding. The *per-class consistency* loss term first computes the degree of consistency between the semantic heat map edges within each channel of $\mathbf{S}$ and the depth map edges at class level. Finally, this quantity is summed up over all classes:

$$L_p = \sum_{c=1}^{C} mean(\{|\frac{dS}{dx}(i,j,c)| \cdot e^{-|\frac{dD}{dx}(i,j)|}\}_{\substack{1 \le i \le N_1, \\ 1 \le j \le N_2}}) + mean(\{|\frac{dS}{dy}(i,j,c)| \cdot e^{-|\frac{dD}{dy}(i,j)|}\}_{\substack{1 \le i \le N_1, \\ 1 \le j \le N_2}}) \quad (1)$$

This formula is based on a simple dissimilarity metric of the form:

$$w(a,b) = |a| \cdot e^{-|b|}. \quad (2)$$

In our case, a/b is semantic/depth edge intensity at a specific image pixel, respectively. In regions with intense depth edges $\lim_{b \to \pm\infty} e^{-|b|} = 0$, therefore semantic edges are not discouraged if depth edges are present. However, in absence of depth edges $\lim_{b \to 0} e^{-|b|} = 1$, so the per-class consistency loss is positive ($w(a,b) > 0$) if, simultaneously, the spatially coinciding semantic map region has a non-zero gradient ($a > 0$).

3.3. Holistic consistency loss

The proposed holistic consistency loss term L_h is a simplification of the per-class consistency loss, where, instead of class-wise edge comparison, the global predicted semantic boundaries are used. These boundaries can be formed by choosing the maximum value among all class semantic edges, for each pixel:

$$L_h = mean(\{S'_x(i,j) \cdot e^{-|\frac{dD}{dx}(i,j)|}\}_{\substack{1 \le i \le N_1, \\ 1 \le j \le N_2}}) + mean(\{S'_y(i,j) \cdot e^{-|\frac{dD}{dy}(i,j)|}\}_{\substack{1 \le i \le N_1, \\ 1 \le j \le N_2}}), \quad (3)$$

where $\mathbf{S}'_k = \{max(|\frac{d\mathbf{S}}{dk}(i,j)|)\}_{\substack{1 \le i \le N_1, \\ 1 \le j \le N_2}}$.

Evidently, the proposed holistic consistency variant is more computationally efficient.

4. EXPERIMENTAL EVALUATION

To assess the proposed L_h method, a popular U-net [6] with a ResNet-50 backbone CNN, pretrained on consecutive video frame pairs, was selected as the depth estimation neural branch, since it does not rely on stereoscopic 3D input. A popular, fast semantic segmentation CNN was selected as the main neural branch [20], serving as our baseline algorithm. The "road01" subset of the Apolloscape dataset [19] was employed for evaluation purposes; to the best of our knowledge, it is the only publicly available and sufficiently large video segmentation dataset (video data are required by [6]). An example video frame is shown in Figure 1. In all cases the original image resolution of 3384×2710 was reduced to 832×256, during both training and testing.

The proposed regularizer L_h was evaluated by comparing the semantic segmentation performance of [20], trained with L_h, against the baseline [20], trained without L_h. Additionally, three properly adapted, recent, competing methods were also evaluated: [17], [12] and [16]. First, the state-of-the-art multitask architecture described in [17] was implemented and the hyperparameter values reported in the paper were further tuned by us for fair comparison. Alternatively, the consistency loss term proposed in [12], which is reminiscent of ours, was implemented on top of [20], instead of L_h, and evaluated under two setups: a) training with a pretrained depth

(a) (b) (c)

Fig. 1: (a) Input image [19], (b) estimated depth map [6], (c) estimated semantic segmentation map (ours))

Table 1: Evaluation results on the Apolloscape dataset, employing a CNN backbone pretrained on ImageNet for image classification. The baseline semantic segmentation branch is [20] and the depth estimation branch is [6]. Methods reported as "pretrained" use a pretrained depth estimation branch with frozen parameters, while methods reported as "multitask" jointly train the two branches. Reported inference time per video frame is an average over the test set.

Method	Mean IoU	Inference runtime (msec)
Baseline (no depth)	39.557%	6.2
[17] (multitask)	34.318%	6.4
Baseline + [16] (multitask)	37.683%	8.3
Baseline + [12] regularizer (pretrained)	39.610%	6.2
Baseline + [12] regularizer (multitask)	38.153%	9
Baseline + L_h (pretrained, proposed)	**40.597**%	**6.2**

network branch (as in the proposed method), and b) training it under a multitask setting, where the semantic segmentation and the depth estimation tasks are learnt simultaneously (as in the original [12]). Finally, the multitask method [16] (without a colorization decoder) was also implemented and plugged on top of [20]. Note that the original [12] and [16] algorithms utilize stereoscopic 3D images during multitask training, thus requiring special datasets. Therefore, for fair comparison, we adapted them to our setting of self-supervised depth estimation from regular, monocular, RGB video using [6], which is arguably a more difficult task. Vanilla [17] already relies on self-supervised depth estimation from RGB video, as is the case with the proposed method.

The evaluation results are depicted in Table 1. Training with the competing loss term from [12] leads only to marginal performance increases over the baseline, since this term mainly improves estimated depth map accuracy, which in turn may offer better scene geometry insights for semantic segmentation. Moreover, in the cases of [17], [12] and [16], multitask training for joint semantic segmentation and self-supervised depth estimation from video does not improve segmentation performance. If the backbone CNN's complexity is not increased, it is a particularly difficult task for multitask learning to handle, compared to using a pretrained depth estimator, thus leading to underfitting. This highlights the advantage of the proposed method in comparison to competing multitask training approaches, when low model complexity (and, thus, low runtime inference requirements) is an important consideration, as in autonomous systems and robotics applications.

Overall, training with the proposed L_h gives a boost of about 1% in the common mIoU metric over the baseline, while also surpassing [17] and the adapted versions of [12] and [16]. The main advantage of the proposed method is its ability to enhance semantic segmentation performance with zero runtime inference overhead, without requiring a complex CNN model, special ground-truth training data (RGB-D, stereoscopic 3D) or special sensors during deployment. Thus, it is most suited to embedded applications such as autonomous systems, robots, vehicles, etc.

5. CONCLUSIONS

This paper showed that although exploitation of scene geometry information may augment semantic segmentation performance using Convolutional Neural Networks, it is not mandatory for geometry to be known or estimated during actual CNN deployment/inference. A novel regularizer was proposed that penalizes differences between semantic and depth predictions on presumed object boundaries during segmentor training. Neither ground-truth depth maps or special data (e.g., stereoscopic 3D) at the training stage, nor known or estimated depth maps at the inference stage are required. Quantitative evaluation was performed on a public scene parsing video dataset for autonomous driving. The results indicate that depth map utilization only during training, without resorting to a multitask setup which may demand a more complex backbone CNN to avoid underfitting, can be sufficient for increasing accuracy during deployment/inference. The obtained model predicts better semantic maps, with zero increase in computational requirements at the inference stage.

6. REFERENCES

[1] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments," *IEEE Signal Processing Magazine*, vol. 36, no. 1, 2018.

[2] S. Papadopoulos, I. Mademlis, and I. Pitas, "Neural vision-based semantic 3D world modeling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[3] C. Symeonidis, E. Kakaletsis, I. Mademlis, N. Nikolaidis, A. Tefas, and I. Pitas, "Vision-based UAV safe landing exploiting lightweight Deep Neural Networks," in *Proceedings of the International Conference on Image and Graphics Processing (ICIGP)*. 2021, ACM.

[4] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[5] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang, "Learning for disparity estimation through feature constancy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *Proceedings of Advances in neural information processing systems (NIPS)*, 2019.

[7] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter," *The International Journal of Robotics Research*, vol. 37, no. 4-5, 2018.

[8] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[9] Y. Cao, C. Shen, and H. T. Shen, "Exploiting depth from single monocular images for object detection and semantic segmentation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, 2017.

[10] J. Jiao, Y. Wei, Z. Jie, H. Shi, R.WH. Lau, and T.S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] S. Zhu, G. Brazil, and X. Liu, "The edge of depth: Explicit constraints between segmentation and depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] P.Y. Chen, A. H Liu, Y.C. Liu, and Y.C.F Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano, "Geometry meets semantics for semi-supervised monocular depth estimation," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 2018.

[14] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," *arXiv preprint arXiv:2002.12319*, 2020.

[15] H. Jiang, G. Larsson, M. M. G. Shakhnarovich, and E. Learned-Miller, "Self-supervised relative depth learning for urban scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[16] J. Novosel, P. Viswanath, and B. Arsenali, "Boosting semantic segmentation with multi-task self-supervised learning for autonomous driving applications," in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2019.

[17] M. Klingner, A. Bar, and T. Fingscheidt, "Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[18] L. Hoyer, D. Dai, Y. Chen, A. Köring, S. Saha, and L. Van Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," *arXiv preprint arXiv:2012.10782*, 2020.

[19] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[20] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.