Improving Semantic Segmentation through Spatio-Temporal Consistency Learned from Videos

Ankita Pasad[1]* ankitap@ttic.edu
Ariel Gordon[2] gariel@google.com
Tsung-Yi Lin[2] tsungyi@google.com
Anelia Angelova[2] anelia@google.com

[1] Toyota Technological Institute at Chicago
[2] Robotics at Google

Abstract

We leverage unsupervised learning of depth, egomotion, and camera intrinsics to improve the performance of single-image semantic segmentation, by enforcing 3D-geometric and temporal consistency of segmentation masks across video frames. The predicted depth, egomotion, and camera intrinsics are used to provide an additional supervision signal to the segmentation model, significantly enhancing its quality, or, alternatively, reducing the number of labels the segmentation model needs. Our experiments were performed on the ScanNet dataset.

1. Introduction

The computer vision community has seen immense progress in solving a variety of semantic image understanding tasks, such as classification and segmentation. Typically, a deep convolutional network learns to predict labels from pixels, remaining mostly unaware of the geometric and physical constraints that govern the visual world.

Learning from video streams, as opposed to images, offers temporal coherency as a strong cue that can significantly enhance segmentation. These cues are often utilized [12] through dedicated network architectures, capable of both segmenting and correlating objects in time.

3D Multiview consistency is another cue, shown [6, 10] to improve semantic segmentation, both as an additional supervision signal to train a single-frame segmenter, and as an additional signal at multi-frame inference time. However, these methods generally require RGBD inputs.

Recent advances in unsupervised depth and egomotion estimation can bring together temporal continuity and multiview consistency as supervision signals for improving segmentation models. A depth prediction model can transform video sequences to RGBD sequences, correspondences between pixels in neighboring frames can be established, and consistency in segmentation mask across corresponding pixels can be used as a supervision signal. Together with the recently-demonstrated [5] ability to learn the camera intrinsics from unlabeled videos, depth and egomotion prediction networks facilitate adapting techniques that were previously reserved to RGBD input to general unlabeled video.

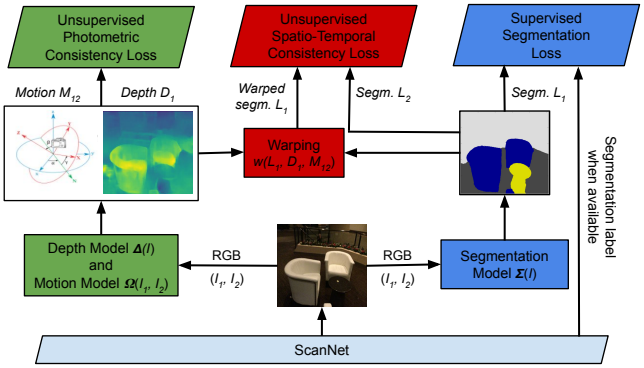

Figure 1: Illustration of the proposed consistency-driven training of the segmentation model.

In this work, we demonstrate the effectiveness of 4D consistency constraints (that is, 3D multiview consistency alongside with temporal continuity) in providing an additional supervision signal for a semantic segmentation task. The latter is trained in a semi-supervised manner, that is, with only a small fraction of the segmentation labels used, whereas the depth, motion and intrinsics models needed for asserting the 4D consistency are trained fully unsupervised. By only using previously-published off-the-shelf networks, we demonstrate that our approach does not require any network architecture tuning, lending itself to future improve-

*Work done while at Robotics at Google.

ments by upgrading the respective models as they become available. The focus of our study can be summarized in three questions:

1. How to construct a training objective that enforces spatio-temporal consistency?
2. How much improvement can the spatio-temporal (4D) consistency constraints provide when training a single-image semantic segmentation model?
3. How many labels can the consistency constraints replace?

The study is performed on the ScanNet [2] dataset.

2. Related Work

The advantages of learning geometric and semantic tasks jointly have been long recognized [3, 13, 15]. Eigen and Fergus [3] are among the first to exploit relations between geometry and semantics by learning them jointly. Their work provides empirical evidence for performance improvement when optimizing depth and semantics jointly. Right-left geometric consistency has been used to improve semantic segmentation in a stereo setting [1, 8]. Optical flow has been used to improve the temporal consistency of segmentation masks [7, 11]. To the best of our knowledge, this work is the first attempt at improving single-image segmentation by employing spatio-temporal consistency using unsupervised depth, egomotion, and camera intrinsics estimates learned from videos.

3. Method

In this section we discuss the formulation of consistency losses that are based on the idea of self-supervision and the design does not assume access to any task specific supervision. We start with a pair of consecutive images from a video, $\mathcal{I}_t$ and $\mathcal{I}_{t+1}$. We have access to the models for depth estimation, camera motion estimation, and semantic segmentation: $\Delta(\cdot)$, $\Omega(\cdot,\cdot)$, and $\Sigma(\cdot)$ respectively. We first estimate the depth $d_t = \Delta(\mathcal{I}_t)$ and the camera motion, i.e., 3D rotation and translation, $M_{t,t+1} = \Omega(\mathcal{I}_t, \mathcal{I}_{t+1})$. Simultaneously, the segmentation model generates the logits masks, $L_t = \Sigma(\mathcal{I}_t)$ and $L_{t+1} = \Sigma(\mathcal{I}_{t+1})$. Using the depth and motion estimates we have a differentiable warping function, $\omega(\cdot,\cdot,\cdot)$ that gives us an estimated transformation function from $\mathcal{I}_t$ to $\mathcal{I}_{t+1}$. We thus have an additional estimate for the logits mask $\hat{L}_{t+1} = \omega(L_t, d_t, M_{t,t+1})$. An overview of the proposed method is presented in fig. 1.

Now, we employ consistency between the propagated logits mask, $\hat{L}_{t+1}$, and the predicted logits mask, L_{t+1}. Note that the proposed loss formulation will hold for the backward consistency constraint between $\hat{L}_t$ and L_t as well, where $\hat{L}_t = \omega(l_{t+1}, d_{t+1}, M_{t+1,t})$.

$$\ell_{L1} = \sum_{x,c} W(x,c)||\hat{L}_{t+1}(x,c) - L_{t+1}(x,c)||_1 \quad (1)$$

where x is the pixel index in 2D space, c is the class index, $W(\cdot,\cdot)$ is the normalized weight for the L1 difference as a function of pixel location and class label. We use a combination of 3 different formulations of the weighing function with the respective weights as 0.2, 0.4, and 0.4.

1. Uniform: Mean of the difference, where $W(x,c)$ is constant across both pixel and class indices in the mask.
2. Label prior: Uniform averaging fails to differentiate between the classes that actually appear in the image from those that don't, whereas it is more reasonable to have a higher penalty for the inconsistencies in the former. Since we do not have the groundtruth labels we use $\hat{y}_{t+1}$ as a belief for the same and set $W(x,\cdot) = \mathbb{1}_{\hat{y}_{t+1}(x)}$, where $\mathbb{1}_{\hat{y}_{t+1}(x)}$ is a one-hot vector of the length same as the number of classes with 1 at index $\hat{y}_{t+1}(x)$, and $\hat{y}_{t+1} = \text{argmax}(L_{t+1})$ along the class index axis.
3. Pixel prior: Here, the weight is constant across different classes while the inconsistencies for edge pixels are penalized more than the others. Here, $W(\cdot,c) = E(\mathcal{I}_{t+1}) \; \forall \; c$, where $E(\mathcal{I}_{t+1})$ is the two-dimensional edge detector output for the image, with one for pixel locations corresponding to the edges.

4. Experiments

4.1. Models

All models trained in our experiments – depth prediction, egomotion prediction, and semantic segmentation – were taken from other publications, using their respective open-sourced code, along with their tuned optimization hyper-parameter settings. This choice allows gauging the quality improvements associated with imposing consistency constraints, as opposed to architectural improvements. By applying the consistency constraints in an architecture-agnostic manner, we leave an open route to further improvements, by simply replacing the comprising models by better ones, as they become available in the literature. For semantic segmentation we use the NAS-FPN [4] as the backbone architecture with the segmentation classifier design as proposed by Kirillov et al. [9]. For the prediction of depth, egomotion, and camera intrinsics we use the models from Ref. [5]. Both the off-the-shelf models we use are recently proposed, strong models attaining the state of the art performance in the respective tasks.

4.2. Dataset

We use ScanNet [2], a dataset of indoor RGB-D video sequences. It consists of 2.5M views across 1500 scans. All the frames in a video sequence are labeled for semantic segmentation masks across 21 classes including a background class. The annotations were obtained by rendering the 3D scans from the sequence of 2D images to get 1500 3D scans. These 3D scans were then manually annotated for segmen-

% supervision	0.1	0.2	0.5	1	2	4
Baseline	39.8	43.2	47.0	49.0	49.1	51.1
With consistency	43.3	46.1	49.4	49.3	51.1	51.1

Table 1: Mean intersection over union (MIOU) scores (%) for semantic segmentation on ScanNet validation set, for models trained with varied fraction of the labels (% supervision), with and without spatio-temporal consistency. No improvement in the MIOU was observed above 4% supervision due to the strong correlations between the ScanNet images.

tation and were projected back to 2D. This is a good dataset for the proof of concept evaluation of our method as it gives a handle to freely control the available supervision to create an artificial limited supervision setting.

4.3. Key Results

The effect of spatio-temporal consistency on the segmentation performance is summarized in Table 1. The fraction of labels used for training is varied from 0.1% to 4%, where the rest of the images were stripped of their labels and only used for imposing spatio-temporal consistency. At or above 4% supervision, we observe a mean intersection over union (MIOU) of 51.1%[1]. For each case, the supervised baseline model is obtained by training the single-image segmentation model on the respective labeled data. The MIOU of the resulting baseline model is summarized in the second row of Table 1. Depth, egomotion, and camera instrinsics models were trained separately, unsupervised, on the entire ScanNet training dataset. Then the consistency loss in Eq. 1 was switched on as an additional supervision signal for the segmentation model, and the latter continued to train, achieving the MIOU values summarized in the third row of Table 1. We notice that the proposed approach gives consistent improvements across the range of supervision. Higher relative improvements are observed at lower supervision.

Additionally, it is meaningful to analyze the effect of consistency as an alternative to direct supervision. Indeed, supervision obtained through consistency constraints can mimic an increase in the number of labels by up to a factor of *four*: The baseline MIOU at 2% supervision (49.1%) matches the MIOU with *only quarter* as much labeled data (49.4%) with the proposed approach.

4.4. Effect on the Rare Labels

Fig. 2 presents the relative improvement in per-class MIOU for the 0.1% baseline when trained with the proposed consistency constraints. We observe that the relative improvement in the MIOU tends to be the highest for the rarest labels, especially those that are only 5% frequent or

[1]This number is in the ballpark of prior RGB-only segmentation benchmarks of MIOU=50.3% [14], however the numbers are not directly comparable since our results were evaluated on the validation set rather than the test set, and because different subsets of the training set were used.

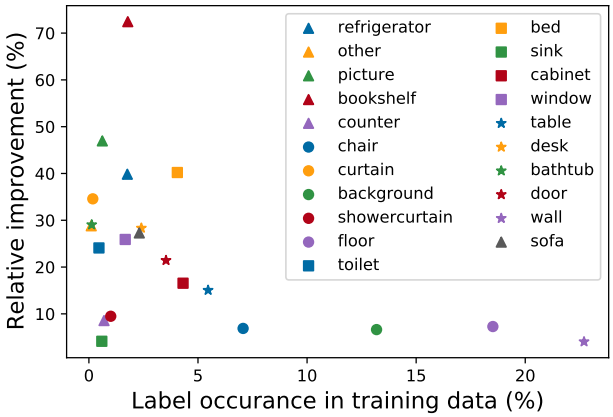

Figure 2: Relative improvement in the class-wise MIOU score as a function of the class frequency in the labeled data.

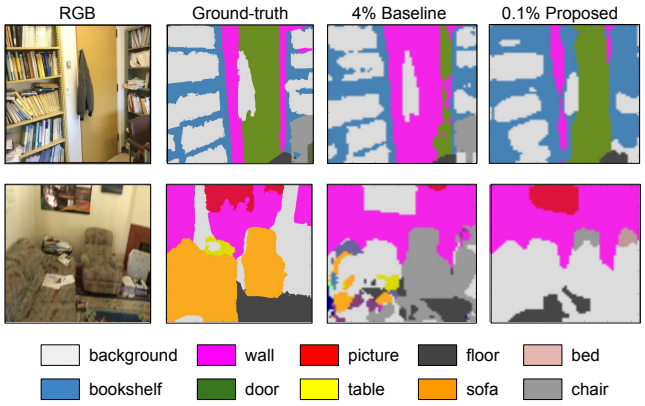

Figure 3: Qualitative comparison of the 4% baseline with the proposed approach trained on 0.1% supervision. Rare labels "door" (top) and "picture" (bottom) are successfully identified by the proposed approach.

lesser. The spatio-temporal consistency constraints effectively augment the training labels to include more instances of each label. This naturally leads to improved performance for especially less frequent labels.

Fig. 3 demonstrates two such examples where the best baseline fails to identify masks for rare labels, "door" and "picture", whereas the proposed consistency-based approach trained with *40 times less direct supervision* is able to accurately label the corresponding pixels. The examples also show that the predictions for the proposed model are smoother across pixels corresponding to a particular object with hardly any fragmentation artifacts.

4.5. Ablation Study of consistency constraints

Table 2 presents the individual contribution of different components of the consistency loss across rows 2 to 5. We can see that each loss components helps individually while the combination of the three performs the best. Additionally

Uniform	Label Prior	Pixel Prior	ℓ_{CE}	MIOU
baseline				39.8
✓				41.7
	✓			41.3
		✓		41.8
✓	✓	✓		**43.3**
			✓	41.4
✓	✓	✓	✓	42.0

Table 2: Effect of different consistency losses (detailed in Sec. 3) on the segmentation performance.

the last two rows of the table are added for reference where the logits are used to generate the pseudo labels. These pseudo labels are then used to train the warped logits using the cross-entropy loss $\ell_{CE} = CE(\hat{L}_{t+1}, \hat{y}_{t+1})$. We see that this straight-forward approach based on the hard decision on predictions is not nearly as good as the proposed averaging method to impose the consistency.

In all the experiments above, the depth, egomotion, and camera intrinsics models supervised the segmentation model via spatio-temporal consistency losses, but not the other way around. Allowing the segmentation model to supervise the other models resulted in no significant improvement in the depth estimation error. While such improvements were observed in prior work [5], they were attributed to the ability of segmentation to identify moving objects. This ability is irrelevant to ScanNet's static scenes. The analyses and the ablation studies presented in this section were done on the 0.1% supervision case.

5. Conclusion

In this work, we used models predicting depth, egomotion, and camera intrinsics, to provide additional supervision to a semantic segmentation model through spatio-temporal consistency constraints. The latter were shown to reduce the need for direct supervision by a factor of up to four. Enhancement in semantic segmentation performance was observed, especially for the less frequent labels. All models were adopted from prior publications, through our approach that is network-architecture-agnostic.

The method proposed in this work can be readily extended to dynamic scenes. Rather than only estimating camera motion, dynamic scenes require the estimation of 3D object motion relative to the scene. It has been previously shown [5] that segmentation can provide a regularization for 3D motion estimation. The consistency losses developed in this study can provide supervision from the depth and motion model to the segmentation model, closing the loop on the three models, depth, 3D motion and segmentation, peer-supervising each other.

References

[1] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2624–2632, 2019.

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5828–5839, 2017.

[3] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE international conference on computer vision, pages 2650–2658, 2015.

[4] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7036–7045, 2019.

[5] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[6] Joris Guerry, Alexandre Boulch, Bertrand Le Saux, Julien Moras, Aurélien Plyer, and David Filliat. Snapnet-r: Consistent 3d multi-view semantic labeling for robotics. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 669–678, 2017.

[7] Junhwa Hur and Stefan Roth. Joint optical flow and temporally consistent semantic segmentation. In European Conference on Computer Vision, pages 163–177. Springer, 2016.

[8] Huaizu Jiang, Deqing Sun, Varun Jampani, Zhaoyang Lv, Erik Learned-Miller, and Jan Kautz. Sense: A shared encoder network for scene-flow estimation. In Proceedings of the IEEE International Conference on Computer Vision, pages 3195–3204, 2019.

[9] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6399–6408, 2019.

[10] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with rgb-d cameras. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 598–605, 2017.

[11] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6819–6828, 2018.

[12] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[13] Trevor Standley, Amir R Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? arXiv preprint arXiv:1905.07553, 2019.

[14] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. International Journal of Computer Vision, pages 1–47, 2019.

[15] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In Proceedings of the European Conference on Computer Vision (ECCV), pages 36–53, 2018.