Composite Learning for Robust and Effective Dense Predictions (Supplementary Material)

Menelaos Kanakis[1] Thomas E. Huang[1] David Bruggemann[1] Fisher Yu[1] Luc Van Gool[1,2]

[1]ETH Zürich [2]KU Leuven

A. Monocular Depth Estimation

Joint optimization on identical dataset subsets Table S.1 presents the performance of the monocular depth estimation single-task baseline and the best performing self-supervised task, DenseCL. While in the main paper (Sec. 4.1, Table 1) the self-supervised objective had access to the entire dataset, in Table S.1 both objectives use the same subset for optimization. Consistent improvements across all dataset splits are still observed.

Table S.1. Monocular depth estimation performance in RMSE on NYUD-v2. Both supervised and self-supervised objectives use identical splits. CompL denotes the addition of the best performing self-supervised objective, DenseCL, and yields consistent improvements.

CompL	Dataset Size				
	5%	10%	20%	50%	100%
	0.8871	0.8120	0.7471	0.6655	0.6223
✓	**0.8840**	**0.8080**	**0.7305**	**0.6508**	**0.5990**

B. Semantic Segmentation

Joint optimization on identical dataset subsets Table S.2 presents the performance of the semantic segmentation single-task baseline and the best performing self-supervised task DenseCL. Similar to Table S.1, both objectives use the same subset for optimization. Consistent improvements across all dataset splits are again observed.

Robustness to zero-shot dataset transfer In Sec. 4.2 , we additionally investigated the generalization capabilities of CompL to a new and unseen dataset. Table S.3 presents the performance of the BDD100K experiments from Fig. 6.

We additionally evaluate how the models trained on PASCAL VOC from Table 2 ("Semseg" and "Semseg + *Task name*") perform without re-training on COCO [7] on the same classes. As seen in Table S.4 and Fig. S.1, joint training with the contrastive methods consistently outperform across all percentage splits, with the lower labeled percentages observing the biggest improvement.

Table S.2. Semantic segmentation performance in mIoU on PASCAL VOC. Both supervised and self-supervised objectives use identical splits. CompL denotes the addition of the best performing self-supervised objective, DenseCL, and yields consistent improvements.

CompL	Dataset Size						
	1%	2%	5%	10%	20%	50%	100%
	30.82	37.66	49.95	55.17	61.30	67.38	70.42
✓	**31.59**	**38.85**	**50.87**	**56.45**	**61.92**	**68.06**	**71.15**

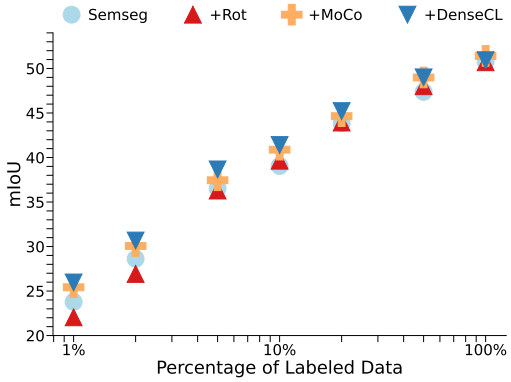


Figure S.1. Performance of semantic segmentation in mIoU trained on PASCAL VOC and evaluated on COCO. The local contrastive loss of DenseCL provides consistent robustness improvements.

C. Multi-Task Model (Semseg and Depth)

Joint optimization In Table 4 of the main paper, we presented the performance of the baseline multi-task model (Depth + Semseg), and the model trained jointly with DenseCL (Depth + Semseg + DenseCL). For ease in comparison between the different models, Fig. S.2 additionally visualizes the results. Training under CompL enhances the performance of both Semseg and Depth, with Depth observing a noticeable gain over Semseg in low data regimes. As discussed in the main paper, this can be attributed to the DenseCL hyperparameters being optimized directly for the improvement of Depth. Furthermore, more advanced loss

Table S.3. Performance of semantic segmentation in mIoU trained on PASCAL VOC and evaluated on BDD100K. The local contrastive loss of DenseCL provides significant robustness improvements.

Model	Labeled Data						
	1%	2%	5%	10%	20%	50%	100%
Semseg	8.18	8.95	10.16	11.18	13.45	17.95	19.51
Semseg + Rot	9.41	8.42	10.71	12.25	13.00	18.00	17.45
Semseg + MoCo	8.56	9.28	11.8	12.28	14.56	20.79	20.45
Semseg + DenseCL	**10.36**	**10.90**	**15.30**	**17.71**	**20.62**	**23.20**	**22.03**

Table S.4. Performance of semantic segmentation in mIoU trained on PASCAL VOC and evaluated on COCO. The local contrastive loss of DenseCL provides consistent robustness improvements.

Model	Labeled Data						
	1%	2%	5%	10%	20%	50%	100%
Semseg	23.78	28.62	36.53	39.05	43.85	47.37	50.76
Semseg + Rot	22.05	26.92	36.29	39.64	43.93	48.01	50.70
Semseg + MoCo	25.42	30.06	37.44	40.88	44.64	48.99	**51.41**
Semseg + DenseCL	**25.96**	**30.67**	**38.66**	**41.40**	**45.21**	**49.00**	50.93

balancing schemes [2] could yield a redistribution of the performance gains, however, such investigation is beyond the scope of our work.

D. Experiment Details

D.1. Codebase

In this work, we base our experiments on the VISion library for state-of-the-art Self-Supervised Learning (VISSL) [3], released under the MIT License. VISSL includes implementations of self-supervised methods, and was adapted to enable for the joint optimization of the existing algorithms with supervised methods (semantic segmentation, monocular depth estimation, and boundary detection). The code will be made publicly available upon publication to spark further research in Composite Learning (CompL).

D.2. Technical details

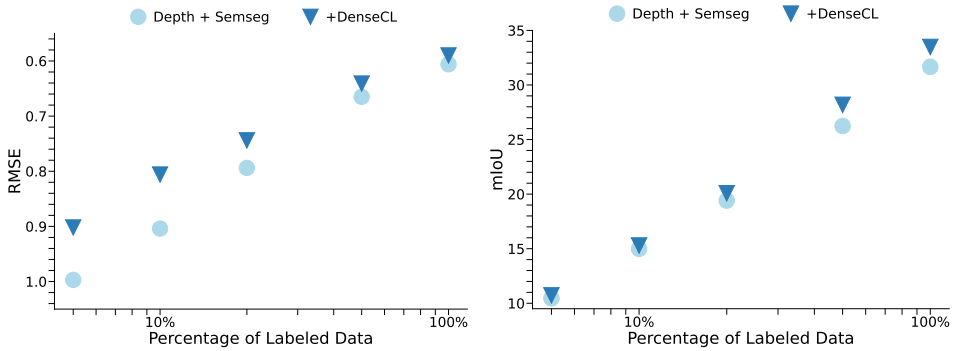
All experiments were conducted in our internal cluster using single V-100 GPUs. Due to the considerable costs associated with multiple runs (beyond our compute infrastructure capabilities), we run all experiments with a random seed of 1, the default setting of VISSL. We provide additional details about different aspects that affect the self-supervised methods below:

Memory bank MoCo [5] and DenceCL [11] utilized a memory bank to enlarge the number of negative samples observed during training, while keeping a tractable batch size. Specifically, both methods use a memory bank of size 65,536. All the datasets we used in our study are of a smaller magnitude compared to that memory bank, e.g. 10,582 and 795 for PASCAL VOC 2012 (aug.) [4] and NYUD-v2 [9], respectively. We therefore set the memory bank to have the same size as the training dataset, yielding a single positive per sample, and therefore allowing for the direct use of the InfoNCE loss [8]. A larger memory bank can also be used, however the contrastive loss would need to be adapted to account for multiple positives [6].

Image cropping We use nearly identical augmentations to those proposed in MoCo v2 [1] for the self-supervised methods of [5, 11], but found it beneficial to modify image cropping. In most classification datasets, each image is comprised of a single object, and thus low overlapping crops can still include the same object. In dense tasks such as semantic segmentation, low overlapping crops can contain different objects (Fig. S.3). We follow the practice of [10] and find a constant crop size and distance between the two patches for each task. We empirically find that square crops of size 384 with a distance of 32 pixels on both axis works best for semantic segmentation, crops of size 283×373 (to maintain input size ratio) with a distance of 8 pixels worked best for depth, and square crops of size 320 with a distance of 4 pixels worked best for boundary estimation.

DenceCL global vs local contrastive DenseCL, as discussed in Sec. 3.2 of the main paper, includes a global and a local contrastive term. The importance of the local contrastive term is weighed by a constant parameter. The original paper found that 0.7 for local contrastive and 0.3 for global contrastive performed best for detection, but used 0.5 to strike a balance between the downstream performance on detection and classification. In our study, we also found 0.7 for local contrastive yields the best performance, and as such, used it for all DenseCL experiments.

Hyperparameter λ During training, the auxiliary loss is



(a) Depth (b) Semseg

Figure S.2. Performance of (a) monocular depth estimation (Depth) and (b) semantic segmentation (Semseg) on NYUD-v2 for their multi-task model. The multi-task model combined with CompL yields consistent improvements in both tasks.

(a) Input image

(b) Blue crop

(c) Purple crop

Figure S.3. Low overlapping crops can be semantically different. This is more apparent in dense prediction datasets where multiple objects can be present in each image.

scaled by the hyperparameter λ, weighting the contribution of the auxiliary self-supervised task. The hyperparameter λ was selected by performing a logarithmic grid search, as commonly done in MTL literature, chosen from the set $\{0.05, 0.1, 0.2, 0.5, 1.0\}$. We found the performance of the models to be consistent when λ is in the range of 0.1 to 0.5, as seen in Table S.5. The performance quickly degrades for values an order of magnitude larger as the model prioritizes the auxiliary task over the target task, while smaller values converge to the baseline performance.

Table S.5. Ablation of the λ parameter for the semantic segmentation model trained jointly with DenseCL. The model yields comparable performance for all three values.

λ	Labeled Data	
	10%	50%
0.1	57.21	68.64
0.2	**57.33**	**68.81**
0.5	57.27	68.79

References

[1] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[2] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

[3] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. https://github.com/facebookresearch/vissl, 2021.

[4] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[9] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[10] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *NeurIPS*, 2020.

[11] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *CVPR*, 2021.