Self-Supervised Learning via Conditional Motion Propagation Supplementary Materials

Xiaohang Zhan[1], Xingang Pan[1], Ziwei Liu[1], Dahua Lin[1], and Chen Change Loy[2]

[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Nanyang Technological University

[1]{zx017, px117, zwliu, dhlin}@ie.cuhk.edu.hk
[2]ccloy@ntu.edu.sg

A. Network Configurations

Here we illustrate the detailed configurations of the network, taking ResNet-50 for example. As shown in Figure 1, the image encoder contains the backbone network and an additional convolution layer to encode images into features with 256 channels. To keep spatial structures in embedded features, we set the total stride in ResNet-50 to 8, and dilations to 2 and 4 for "conv4_x" and "conv5_x", the last two residual groups defined in [2].

B. Fine-tuning Details

VOC2012 Semantic Segmentation (AlexNet). Following previous works, we fine-tune the pre-trained weights on AlexNet for PASCAL VOC 2012 semantic segmentation task with FCN-32s [5] as the head. We remove the additional convolution layer of our image encoder, and fine-tune all the layers. The initial learning rate is 0.01 and it is decayed by 10 times at $30K$, $48K$, $60K$ iterations. The total iteration is $66K$.

VOC2012 Semantic Segmentation (ResNet-50). We fine-tune the ResNet-50 CMP model for $33K$ iterations with an initial learning rate of 0.01, with the polynomial learning rate decay strategy (power: 0.9). All the experiments including baselines, upper bound and our method use the same hyper-parameters.

COCO Instance Segmentation (ResNet-50). We construct new baselines and the upper bound for self-supervised learning on COCO Instance Segmentation. We use ResNet-50 as the backbone and Mask R-CNN [1] with FPN [4] as the head. We use the same hyper-parameters across all the experiments, including an initial learning rate of 0.02, learning rate decaying by 10 times at epoch 10 and 15, and the total epoch is 16. Those hyper-parameters are expected to be fixed for future self-supervised learning studies.

LIP Human Parsing (ResNet-50). We perform a comparison on the validation sets of two sub-tasks, including LIP Single-Person Parsing and LIP Multi-Person Parsing. The fine-tuning epochs are respectively 50 and 120 for these two tasks. The initial learning rate is 0.01, and the learning rate decay strategy is polynomial (power: 0.9). The hyper-parameters are kept the same across all the experiments.

C. Evaluation on Detection

We additionally perform experiments with VGG-16 to compare with a recent multi-task based self-supervised learning method [6] which achieves state-of-the-art with VGG on PASCAL VOC 2007 detection task. We use the released pre-trained model of Wang *et al*. [6] for detection and segmentation evaluation. The evaluating experiments are conducted in the same circumstances. As shown in Table 1, CMP does better in segmentation tasks than detection tasks, since CMP focuses on learning spatial structural representations.

Table 1. Evaluation on VOC 2007 detection and VOC 2012 segmentation. Comparison with Wang *et al*. [6]. For detection of Wang *et al*. [6], 63.2% is reported and 57.0% is reproduced.

	Det. (mAP)	Seg. (mIoU)
ImageNet [3]	67.3	64.1
Random	39.7	35.0
Wang *et al*. [6]	63.2 (57.0)	54.0
CMP	56.8	57.6

D. Visualizations

Testing with Noisy Guidance. To better understand CMP's ability of learning kinematic properties. We deliberately give noisy guidance in testing. As shown in Figure 2, (a)

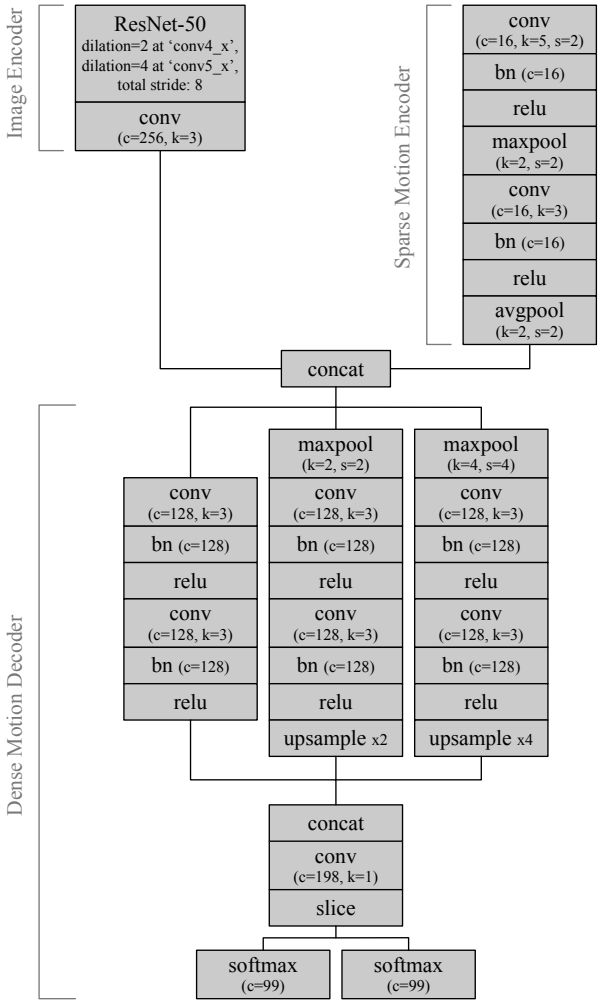

Figure 1. Network configurations, taking ResNet-50 for example. Notations "conv4_x" and "conv5_x" are the last two residual groups defined in [2]. Parameters c, k and s stand for the number of output channels, kernel size and stride.

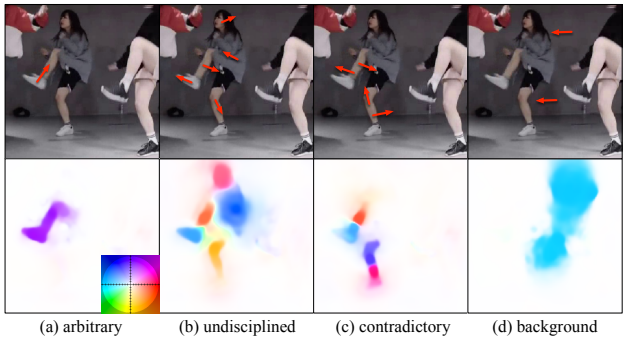

Figure 2. Noisy motion guidance.

Given arbitrary guidance on a single point, rigidity awareness and physical feasibility still hold. (b) Given a group of undisciplined guidance vectors, *i.e.*, given random guidance vectors on different parts, these characteristics hold locally. The global kinematic coherent does not hold expectably, because the CMP model faithfully follows the given guidance, rather than over-fits the image to produce a plausible result. (c) Given contradictory guidance, *i.e.*, given two guidance vectors in different directions on a rigid part, the rigidity awareness does not hold anymore. (d) Given outlying guidance on background, the motions are propagated within the background, while the foreground objects' optical flows are not affected.

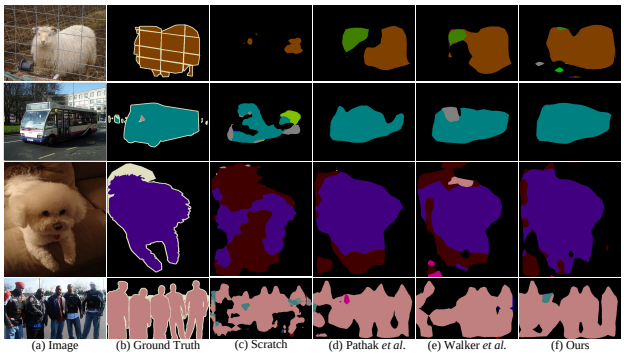

Figure 3. Visual improvements on the validation set of VOC2012 (AlexNet).

Target Tasks. For the fine-tuning tasks on semantic segmentation and human parsing, we show the visual comparisons between our method and baselines in Figure 3 and Figure 4, corresponding to PASCAL VOC 2012 and LIP datasets respectively. When using our CMP pre-trained models, the fine-tuning results are more accurate and spatially coherent. For example, as the first three rows of Fig.3 show, baseline methods misclassify some parts of the sheep, bus, and dog, while our method produces spatially accurate and coherent results. It is due to the kinematically-sound representations learned from CMP.

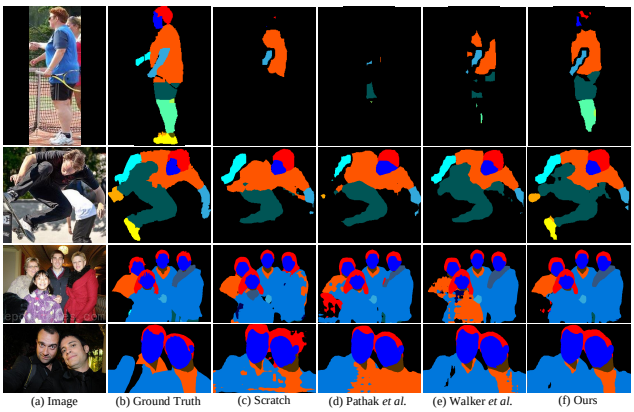

Figure 4. Visual improvements on the validation sets of LIP single-person and multi-person tasks (ResNet-50).

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*. IEEE, 2017. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 1, 2

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[4] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[6] Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for selfsupervised visual representation learning. In *ICCV*, 2017. 1