We thank the reviewers for their time and insightful comments. We address the specific concerns and questions below.

R1 R3 Table 1, 4: Sorry if it was not very clear in the paper. SimCLR with 65.7% in Table 1 is the original SimCLR model with 2-layer MLP in the decoder. In Table 4, we find that by adding an extra 2-layer MLP (4 layers in total) in the decoder can further boost the performance to 67.1%. The 1.4% accuracy gap is comparing 4-layer MLP versus 2-layer MLP, not 2-layer MLP versus no MLP. Therefore our conclusion does not contradict the SimCLR or MoCo papers. Meanwhile, adding two more MLP layers can also help boost performance of GIM: 62.3% (Row 5) vs 60.9% (Row 4) and LoCo: 67.5% (we ran this in addition) vs 66.2%. Since the gain of increasing MLP depth is orthogonal to our paper, we keept the decoder MLP depth as 2 (explained in L264). Note that LoCo contains 1 more conv block in the decoder, but we find that adding conv block into SimCLR does not help (Table 4 Row 3). Hence we believe the comparison is fair.

R1 Results worse than original SimCLR paper: Our results in the main paper are trained with 100 epochs for faster experimentation and the performance is not saturated. The results match Fig.9 in the SimCLR paper which shows performance also with 100 epochs. We also include results trained with 800 epochs in the Supp. Material. SimCLR and LoCo are 69.8% and 69.5% respectively, which match the 69.3% results in the original SimCLR paper.

R1 Transfer learning not studied: Although we did not include CIFAR or Flowers in the paper, we use the more challenging object detection and instance segmentation tasks on COCO and Cityscapes as our transfer learning setting.

R1 Clarity: 1. “we evaluate ...”: As stated “Following [56, 52, 2, 32, 21]”, the encoder is frozen. 2. “"as they do not ...”: Thanks for pointing this out, we intend to say that the intermediate activations can be released when the local module finish its calculation. 3. “"the classification task is ...”: Sorry for the vague expression. We were trying to describe the contribution made by SimCLR. We will revise these sentences.

R1 increasing decoder depth: We believe properly increasing the decoder depth is non-trivial, as adding conv blocks into SimCLR decoder will not help, and simply increasing MLP depth does not help reduce the gap between end-to-end and local learning. We think this is a meaningful finding.

R3 Benefits of sharing features between stages: Our goal is to bridge the gap from two different aspects, sharing stage and making decoder deeper by adding convolution. In fact, by only sharing, we can achieve 4 points gain and achieve 64.9% compared to GIM 60.9%. We will include this in the next version. Only adding convolution into the decoder can get 65.2%. The performance eventually achieves 66.2% by combining both improvements.

R3 Performance of random init: Although the network is randomly initialized, the whole network (including the backbone) is trained with supervised learning. Similar observations can be found in [3].

R3 SyncBN: During the contrastive learning phase, all models including SimCLR and GIM are equipped with SyncBN in our implementation. For downstream tasks, all models use the same head (L208-210), thus SimCLR and GIM also use SyncBN. When finetuning, all models, including the random initialized one, use SyncBN (L227-228).

R4 Biologically plausibility: We thank the reviewer for bringing up this interesting point and we will add more discussion to our paper. Although our algorithm still relies on backprop within a module that consists of multiple layers, a layer in a CNN may not strictly map to a “layer” in the brain. In fact, a few residual blocks grouped together are found to roughly correspond to different regions in the visual cortex (e.g. V1, V2, IT) [1]. Local forward pass and backward pass could be seen as rough approximation of what is being done in the brain using recurrent computation [2]. But how to make it more biologically plausible on a local level is a very exciting future direction!

R4 Computational efficiency: We have not experimented with sequential greedy training and it would be an interesting future step. However, as hypothesized in the paper, one potential drawback of sequential training in GIM is that the lower unit becomes “unaware” of the computation of the upper unit and may not provide the most useful representation. 2.76× memory saving is still substantial. It can make a sizeable difference in memory intensive tasks such as semantic segmentation, where we can hardly fit a single example per GPU.

R4 Clarity: 1. “we are also ...”: we believe performing transfer learning on complex downstream tasks like instance segmentation is essential and valuable to evaluate the quality and transferability of the features learned by contrastive learning. We use the word “the first” as there is no other result that can be referred when we submitted the paper. 2. “Meanwhile, it can ...”: For end-to-end learning, lower layers are required to wait for the gradients from upper layers, and they might be computed in another machine in model parallelism. In local learning, the dependency can be removed and the waiting time can be less. 3. “Importantly, by having ...”: The shared overlapped stage can be treated as part of the decoder for the lower local unit. We believe in this way we effectively make decoders deeper without introducing extra cost in the forward pass (as the computation can be shared for the upper local unit). We will revise these sentences.

R4 Prior work: Thanks for pointing it out, we will compare with Feedback Alignment in the next version.

R5 Advantages of this paper: The goal of this paper is to propose a simple and effective local learning algorithm that can perform as good as SimCLR. We believe the properties of local learning, including lower peak memory footprint and biological plausibility, are very relevant and interesting to the NeurIPS community.

[1] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *bioRxiv, 10.1101/2020.06.16.155556*, 2020

[2] Q. Liao, and T. Poggio. Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. *arXiv preprint 1604.03640*, 2016.

[3] K. He, R. Girshick, and P. Dollár. Rethinking ImageNet Pre-training. *arXiv preprint 1811.08883*, 2018.