Supplementary Material of An Empirical Study of the Collapsing Problem in Semi-Supervised 2D Human Pose Estimation

A. Other Attempts to Avoid Collapsing

The *standard consistency-based method* (Described in Section 3.1 and 3.2) suffers from collapsing problem in 2D pose estimation. In this section, we also investigate the effects of other factors like the number of labeled examples, and unsupervised loss weight on the collapsing problem.

The following experiments show that adjusting these factors can *Not* fully solve the collapse problem. In all experiments, the response to the unlabeled sample always has a downward trend, and the final model accuracy is lower than the initial supervised model. It is worth noting that the augmentation parameters η and η' are sampled in the same distribution (See Section 3.1) in these experiments, and our easy-hard augmentation strategy is not involved.

A.1. Decrease Unsupervised Loss Weight λ

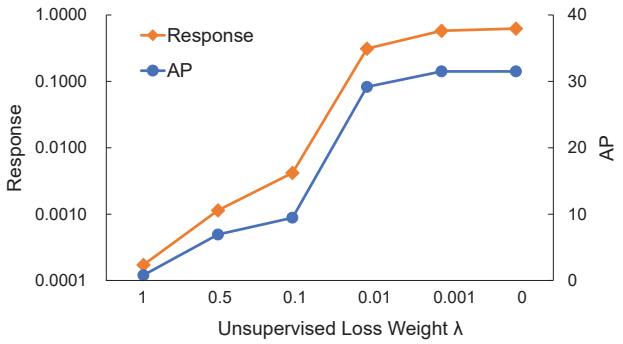

Figure 9. Effect of unsupervised loss weight λ. With small consistency coefficient, the response for unlabeled examples still decreases but in a relatively slow speed.

The coefficient λ controls the importance of consistency loss L_u. One possible opinion is that too large weight λ leads to the instability in training process. The effect of decreasing λ is shown in Figure 9. $\lambda = 1$ is the default setting and $\lambda = 0$ means that only supervised loss L_s is used. All of these are trained with the same number of epochs. The result indicates that lower coefficient slightly reduce the degree of degradation but the degradation still happens. With different coefficient values, the response and AP are always worse than the supervised model.

A.2. Increase Labeled Examples

Labels	Response	AP	Supervised AP
COCO 1K	0.0002	0.8	31.5
COCO 5K	0.0023	13.0	46.4
COCO 10K	0.0105	26.2	51.1
COCO 150K (Full)	0.0122	46.8	67.1

Table 9. Increasing the labeled examples do Not address the collapsing problem. "Supervised AP" represents the supervised model using only labeled examples. Even with sufficient labels, the SSL training still degrades the performance compared to supervised model.

Another question is whether too few labels causes the collapsing. As is shown in Table 9, with the increase of labels, the validation performance has improved (From 0.8% to 46.8%). However, It did not fully solve the collapsing problem, as the response level is still low and performance is degraded. The results indicate that the even under a large number of labeled samples, consistency loss still drives the network to generate low-response heatmap for unlabeled examples, and finally degrade the generalization performance of model. In contrast, our method can significantly improve the performance regardless of the number of labeled examples.

B. Dual Networks

More details about dual networks, including the algorithm flow, the motivation and advantages, are provided in this part. In dual networks learning, we jointly learn two networks f_θ and f_ξ. We first train them separately on labeled images from different initialization. Then we jointly train them on both labeled and unlabeled images. The Figure 10 demonstrates the framework of our method (dual network) and Algorithm 1 describes the algorithm process.

In summary, for an unlabeled image. (1) First generate easy augmentation I_e and hard augmentation I_h. (2) Then the easy augmentation is fed into two networks to produce

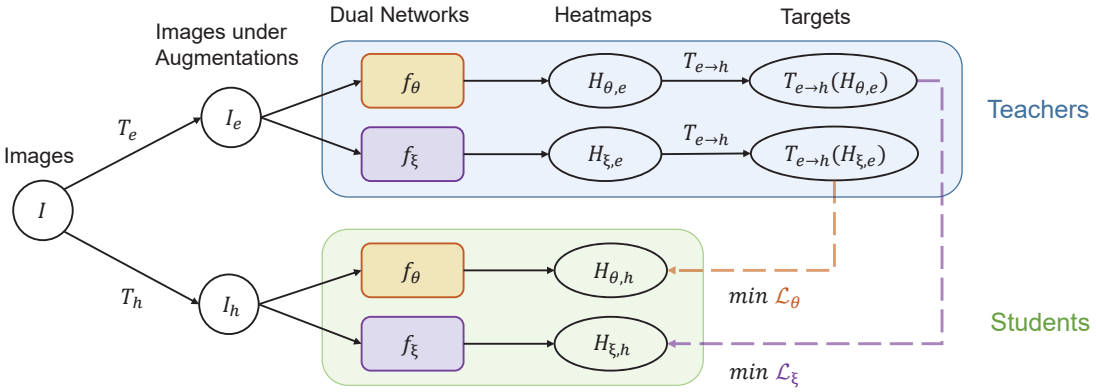

Figure 10. Framework of our Dual Networks Learning. Each of the two networks serves as both a teacher and a student. They take easy and hard images when they are teachers and students, respectively.

targets $H_{\theta,e}, H_{\xi,e}$. (3) And the hard augmentation is used to produce student predictions $H_{\theta,h}, H_{\xi,h}$. (4) The target $H_{\theta,e}$ guides the prediction $H_{\xi,h}$ to optimize f_ξ. And the target $H_{\xi,e}$ guides the prediction $H_{\theta,h}$ to update f_θ.

Such a symmetrical structure is very simple, but has several advantages. (1) It ensures that the teacher prediction is statistically more accurate than the student prediction, which avoid the collapsing problem. (2) The teacher and student come from two different and independent networks, which introduces a certain degree of divergence between two models and boosts the final performance [3, 4]. (3) The two networks can be updated and improved at the same time, and no network will become the bottleneck. In fact, the accuracy of two networks are very similar after training if they use the same network structure.

Algorithm 1 Dual Networks Learning

Input: $\mathcal{L} = \{(\mathbf{I}_l, \mathbf{H}_l)\}_{l=1}^N$: Batch of labeled data.
Input: $\mathcal{U} = \{\mathbf{I}_u\}_{u=1}^M$: Batch of unlabeled data.
Input: θ, ξ: Pre-trained model parameters
Output: θ, ξ: Updated model parameters

$L_s = 0, L_\theta = 0, L_\xi = 0$
for each $(\mathbf{I}_l, \mathbf{H}_l) \in \mathcal{L}$ **do**
  Calculate supervised loss
  $L_s = L_s + ||f(\mathbf{I}_e, \theta) - \mathbf{H}_e||^2 + ||f(\mathbf{I}_e, \xi) - \mathbf{H}_e||^2,$
end for
for each $\mathbf{I}_u \in \mathcal{U}$ **do**
  Randomly sample augmentations η_e and η_h
  $\mathbf{I}_e = T(\mathbf{I}_u, \eta_e),\ \mathbf{I}_h = T(\mathbf{I}_u, \eta_h)$
  Compute teacher predictions $\mathbf{H}_{\theta,e}, \mathbf{H}_{\xi,e}$ by
  $\mathbf{H}_{\theta,e} = f(\mathbf{I}_e, \theta),\ \mathbf{H}_{\xi,e} = f(\mathbf{I}_e, \xi)$
  Generate targets $T_{e\to h}(\mathbf{H}_{\theta,e}),\ T_{e\to h}(\mathbf{H}_{\xi,e})$
  Compute student predictions $\mathbf{H}_{\theta,h}, \mathbf{H}_{\xi,h}$ by
  $\mathbf{H}_{\theta,h} = f(\mathbf{I}_h, \theta),\ \mathbf{H}_{\xi,h} = f(\mathbf{I}_h, \xi)$
  Calculate unsupervised loss L_θ and L_ξ by
  $L_\theta = L_\theta + \|\mathbf{H}_{\theta,h} - T_{e\to h}(\mathbf{H}_{\xi,e})\|^2,$
  $L_\xi = L_\xi + \|\mathbf{H}_{\xi,h} - T_{e\to h}(\mathbf{H}_{\theta,e})\|^2$
end for
$L_{total} = L_s + L_\theta + L_\xi$
update θ, ξ by minimizing L_{total}

C. Additional Experimental Results

C.1. Augmentation Hyper-Parameters

We study the effect of augmentation hyper-parameters in this part. The dual networks (with "easy-hard" augmentation method) is used as training method. The COCO 1K subset is used as labeled set and COCO *TRAIN* is used as unlabeled set. Joint Cutout and RandAugment are only used in hard augmentation. The best value of hyper-parameters is selected and used as the default setting.

Joint Cutout In Joint Cutout, m detected joints are randomly selected to be masked. More masked regions will increase the difficulty of prediction. We conduct experiments to analyze the effect of hyper-parameter m. The experiment (Figure 11) shows that $m = 5$ achieves the best performance. Further increasing m can not boost accuracy, which is maybe because the the remaining image region is too small to have effective information for localization.

Rand Augment We also study the effect of the distortion magnitude [1] in RandAugment . Two augmentation transformations are randomly sampled and applied sequentially. The Figure 12 shows that the optimal magnitudes is 20. For simplicity, the geometric transformations in RandAugment like "Rotate", "TranslateX", etc. are excluded.

Table 10. The results on the *VAL* set of the MPII dataset. Our model is trained on the MPII *TRAIN* with labels and the AIC without labels. We compare to a supervised baseline that is trained only on the MPII *TRAIN*. "Extra Labels" means we use the labels from the AIC dataset.

Method	Backbone	Extra Labels	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	**Total**	Gain
SimpleBaseline [5]	ResNet-101	✗	97.1	94.9	88.1	82.5	87.7	83.2	79.2	88.1	
SimpleBaseline [5]	ResNet-101	✓	97.3	96.2	91.1	86.9	89.3	87.9	83.8	90.8	2.7
Ours	ResNet-101	✗	97.1	96.1	90.4	85.0	89.2	86.1	81.7	**89.8**	↑ **1.7**
SimpleBaseline [5]	ResNet-152	✗	96.6	95.0	89.0	83.4	88.4	84.4	80.3	88.7	
SimpleBaseline [5]	ResNet-152	✓	97.0	96.3	91.3	87.3	89.6	88.5	84.3	91.0	2.3
Ours	ResNet-152	✗	97.2	96.3	91.1	85.9	89.7	87.1	82.9	**90.5**	↑ **1.8**
HRNet [2]	HRNet-W32	✗	97.0	95.7	89.4	85.6	87.7	85.8	82.0	89.5	
HRNet [2]	HRNet-W32	✓	97.4	96.7	92.1	88.4	90.8	88.6	85.0	91.7	2.2
Ours	HRNet-W32	✗	97.4	96.6	91.8	87.5	89.6	87.6	83.8	**91.1**	↑ **1.6**

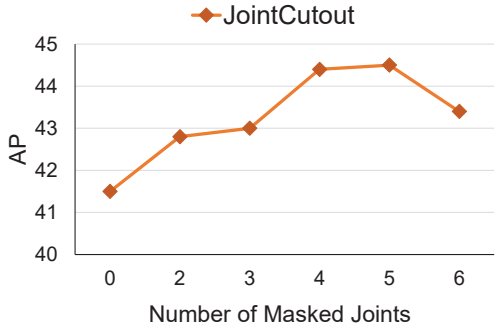

Figure 11. Effect of the number of masked regions in Joint Cutout. The results show the randomly masking some joints in hard augmentation help to improve the performance and $m = 5$ achieves the best accuracy in this setting.

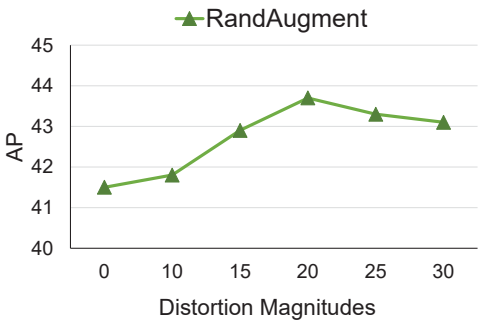

Figure 12. Effect of the distortion magnitudes in RandAugment.

C.2. MPII Validation Dataset

We test our method (Dual) in a more realistic setting where labeled and unlabeled images are from different datasets of MPII and AIC, respectively. Table 10 shows the results on the validation set of MPII. Our approach outperforms the supervised baseline [5] by a large margin on three different backbones. After applying our method to utilize unlabeled data, the error rate reduces by around 15%. It is worth noting that performing supervised learning on the combined MPII and AIC dataset with extra labels only gets slightly better accuracy than our semi-supervised approach which validates the effectiveness of our approach.

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 2

[2] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3

[3] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020. 2

[4] Wei Wang and Zhi-Hua Zhou. Analyzing co-training style algorithms. In Joost N. Kok, Jacek Koronacki, Raomon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *European Conference on Machine Learning*, pages 454–465, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. 2

[5] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*, pages 466–481, 2018. 3