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

Rongchang Xie[1], Chunyu Wang[2], Wenjun Zeng[2], Yizhou Wang[1]

[1]Peking University, [2]Microsoft Research Asia

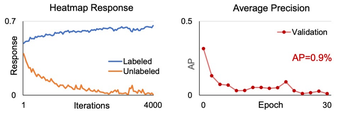
Introduction

- **The Task** Leverage large-scale unlabeled images for training 2D human pose estimator
- **Why this is an important problem?**
  - 2D pose estimation is sensitive to occlusion/self occlusion/viewpoint variations
  - Labeled datasets are small (MPII, COCO, AI Challenger)
  - Poor generalization accuracy (Variations of poses, appearance, background, scale)

- **Status & Challenges**
  - Consistency-based semi-supervised training works well for classification
  - Collapse (predict all as background) when applied to human pose estimation

Supervised

Naïve consistency

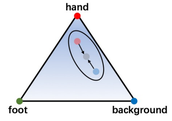


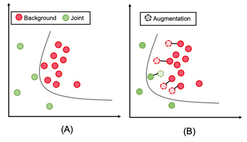
Contributions

- **Analyze the collapsing problem** In this paper, we show that collapsing is caused by the extreme class-imbalance problem in 2D human pose estimation
- **Address the problem by easy-hard augmentation scheme** Based on our analysis, we present a very simple solution by constructing a pair of easy and hard augmentations, and using the easy augmentation to teach the hard one
- **SOTA Results** Our training approach generally applies to the most state-of-the-art methods and achieves several points of improvement by using large-scale unlabeled images

Analysis of the Collapsing Problem

- The standard consistency-based method minimizes the distance between the predictions of the two augmentations (red and blue points). Since many pixels have low response (close to background), few high response pixels (*e.g.*, the red point) tend to be gradually pulled to the background class.

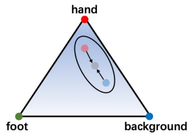




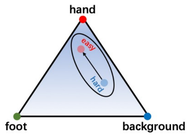
- (A) the decision boundary before SSL.
- (B) the naive consistency regularization moves data and their augmentations (dashed circles) to their middle points. As a result, more data will be close to the decision boundary which pushes the decision boundary to pass through the areas of minor class that is sparse globally

Address Collapsing by Easy-Hard Augmentation

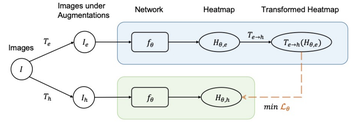
- Instead of drawing two augmentations to their middle point (as in standard consistency learning), **we drive the less accurate predictions which are close to the decision boundary to the direction of more accurate predictions**



Standard consistency learning

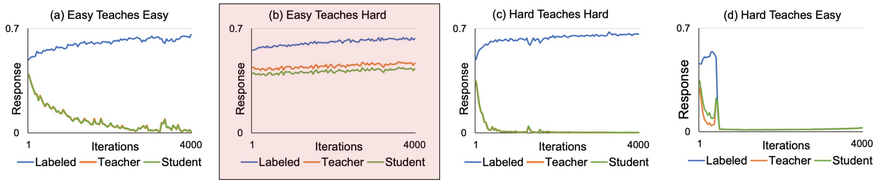


Proposed consistency learning

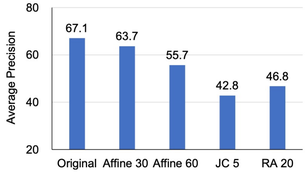


Note that the optimization is unidirectional

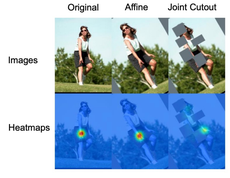
- **The easy hard augmentation significantly stabilizes training**



- **How to construct easy-hard augmentation pair**
  - We determine "easiness" and "hardness" based on its effect on estimation accuracy in testing
  - Selection of easy-hard augmentations affects accuracy
  - The good thing is that the selection seems to generalize across different datasets



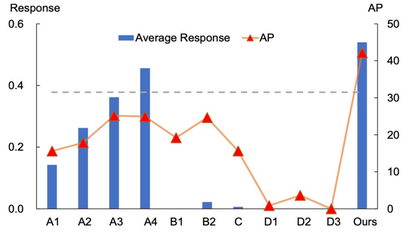
Easy Aug	Hard Aug	Avoid Collapsing
Affine 30	Affine 30	No
Affine 60	Affine 60	No
Original	Affine 60	**Yes**
Affine 30	Affine 60	**Yes**
Affine 60	JC 5	**Yes**



- **Several failed attempts we did in this project**

  - A1-A4: Use confident predictions to teach the network with confidence thresholds of 0.2, 0.4, 0.6 and 0.8, respectively
  - B1-B2: EMA parameters of 0.99 and 0.999
  - C: Class re-weighting method
  - D1-D3: Easy-easy, hard-hard and hard-easy augmentation

  The gray dash lines is the AP of the initial supervised model



Experiments and Results

- **Comparison to other semi-supervised learning methods**

Methods	Aug.	1K	5K	10K	All
Supervised [33]	A	31.5	46.4	51.1	67.1
PseudoPose	A	37.2	50.9	56.0	—
DataDistill [23]	A	37.6	51.6	56.6	—
Ours (Single)	A	38.5	50.5	55.4	—
Ours (Dual)	**A**	**41.5**	**54.8**	**58.7**	—
Ours (Single)	A+JC	42.1	52.3	57.3	—
Ours (Dual)	A+RA	43.7	55.4	59.3	—
Ours (Dual)	A+JC	44.6	55.6	59.6	—

- **Effect of using different network structures for student and teacher models**

Method	Networks of f_θ and f_ξ	1K	5K	10K
Supervised [33]	ResNet18	31.5	46.4	51.1
Supervised [33]	HRNet w48	39.2	57.7	63.7
Ours (Dual)	ResNet18	41.5	54.6	58.6
	ResNet18	41.6	54.9	58.8
Ours (Dual)	HRNet w48	50.9	64.3	67.9
	HRNet w48	51.0	64.2	67.9
Ours (Dual)	ResNet18	**48.7**	**59.4**	**62.5**
	HRNet w48	50.9	62.8	66.8

- **Comparison to the state-of-the-art methods**

Method	Network	Input Size	GFLOPS	#Params	AP	AP0.50	AP0.75	APM	APL	AR
SB [33]	ResNet50	256 × 192	8.9	34.0	70.2	90.9	78.3	67.1	75.9	75.8
SB [33]	ResNet152	256 × 192	15.7	68.6	71.9	91.4	80.1	68.9	77.4	77.5
HRNet [27]	HRNetW48	384 × 288	32.9	63.6	75.5	92.5	83.3	71.9	81.5	80.5
MSPN [19]	ResNet50	384 × 288	58.7	71.9	76.1	93.4	83.8	72.3	81.5	81.6
DARK [36]	HRNetW48	384 × 288	32.9	63.6	76.2	92.5	83.6	72.5	82.4	81.1
UDP [10]	HRNetW48	384 × 288	33.0	63.8	76.5	92.7	84.0	73.0	82.4	81.6
Ours (+SB)	ResNet50	256 × 192	8.9	34.0	**72.3** (↑ 2.1)	91.8	80.5	69.3	77.8	77.7
Ours (+SB)	ResNet152	256 × 192	15.7	68.6	**73.7** (↑ 1.8)	92.1	82.1	71.0	79.0	79.1
Ours (+HRNet)	HRNetW48	384 × 288	32.9	63.6	**76.7** (↑ 1.2)	92.5	84.3	73.5	82.5	81.8
Ours (+DARK)	HRNetW48	384 × 288	32.9	63.6	**77.2** (↑ 1.0)	92.6	84.5	73.9	82.9	82.2

Future Work

- Explore coherence information in videos for unsupervised learning
- Explore possibility of automatically determining teaching/learning signals

Code & References

Our code is released at https://github.com/xierc/Semi_Human_Pose