Learning Human Poses from Actions

Aditya Arun[1], C.V. Jawahar[1], M. Pawan Kumar[2,3]

[1]International Institute of Information Technology, Hyderabad
[2]University of Oxford, [3]The Alan Turing Institute

Aim

- We propose a novel framework to **learn human poses** using **diverse data set**.
- In our setting, diverse data set includes:
  - A few expensively annotated images with ground-truth joint annotations
  - Several cheaply annotated images specifying human action.

Actions and Poses

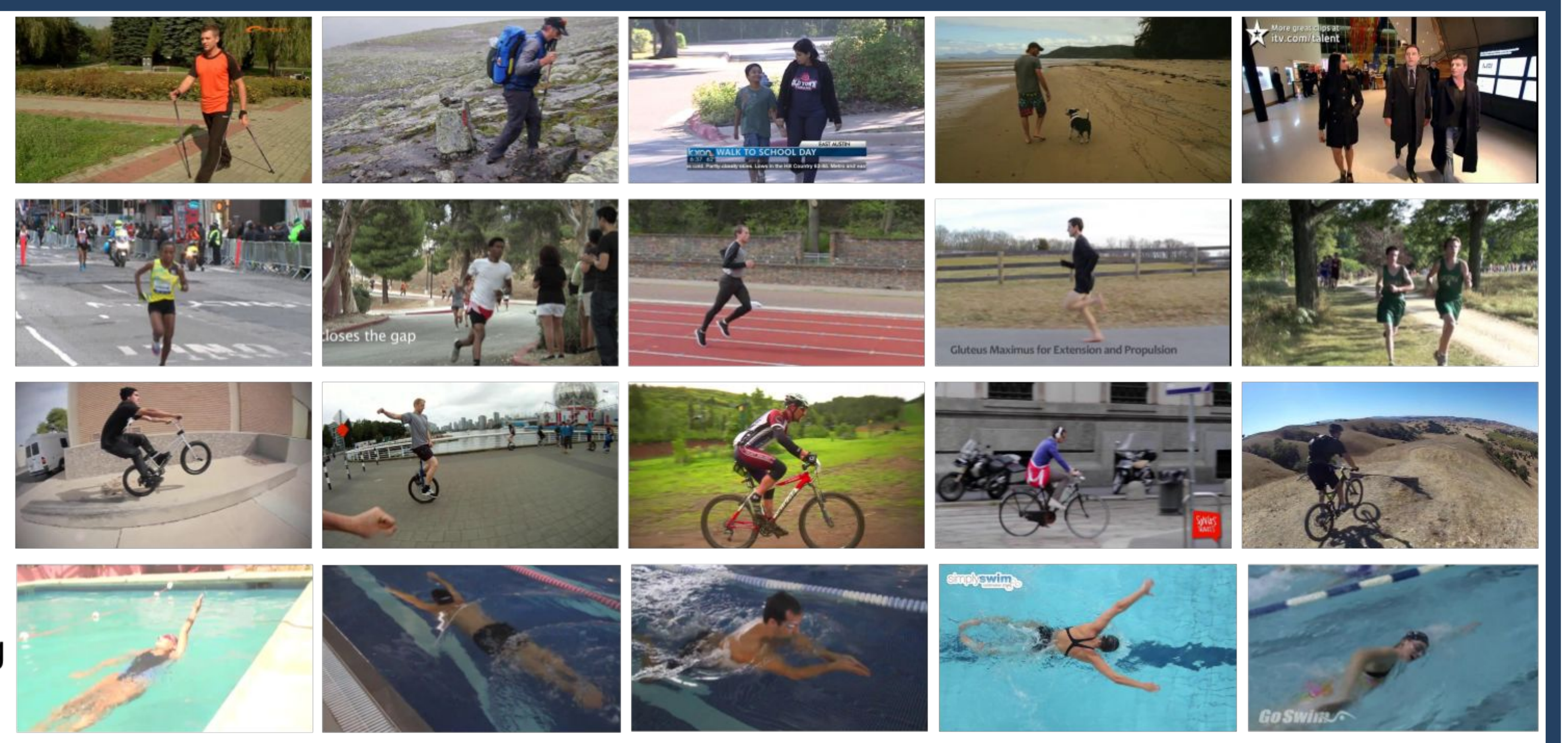


Fig: Images from MPII Human Pose test set for four different action classes.

- Human poses for each action class are sufficiently different.
- Probabilistic formulation necessary due to high intra-class variance.

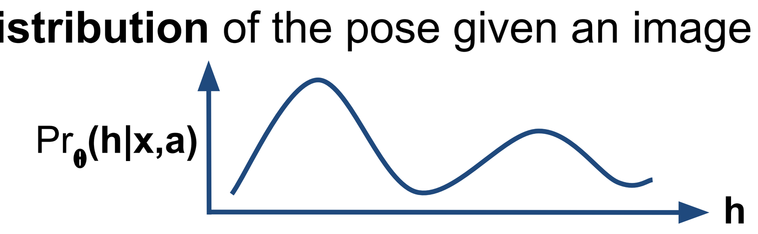
Probabilistic Formulation

Tasks:

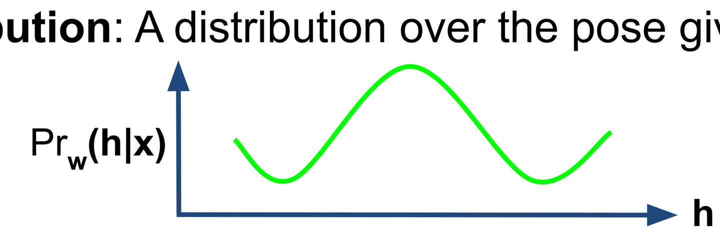
1. During training, model the uncertainty in the pose for every action.
2. During inference, predict the pose given an image.

Two separate distributions for two tasks[1]:

1. A **Conditional Distribution** of the pose given an image and its corresponding action.



2. **Prediction Distribution**: A distribution over the pose given an image.



Ideally, the two distributions should match exactly.

We use a probabilistic network, DISCO Nets[2], to model parameters of these distributions.

DISCO Stacked Hourglass Network

By adding noise filter to the stacked hourglass network[3], we construct the DISCO net.

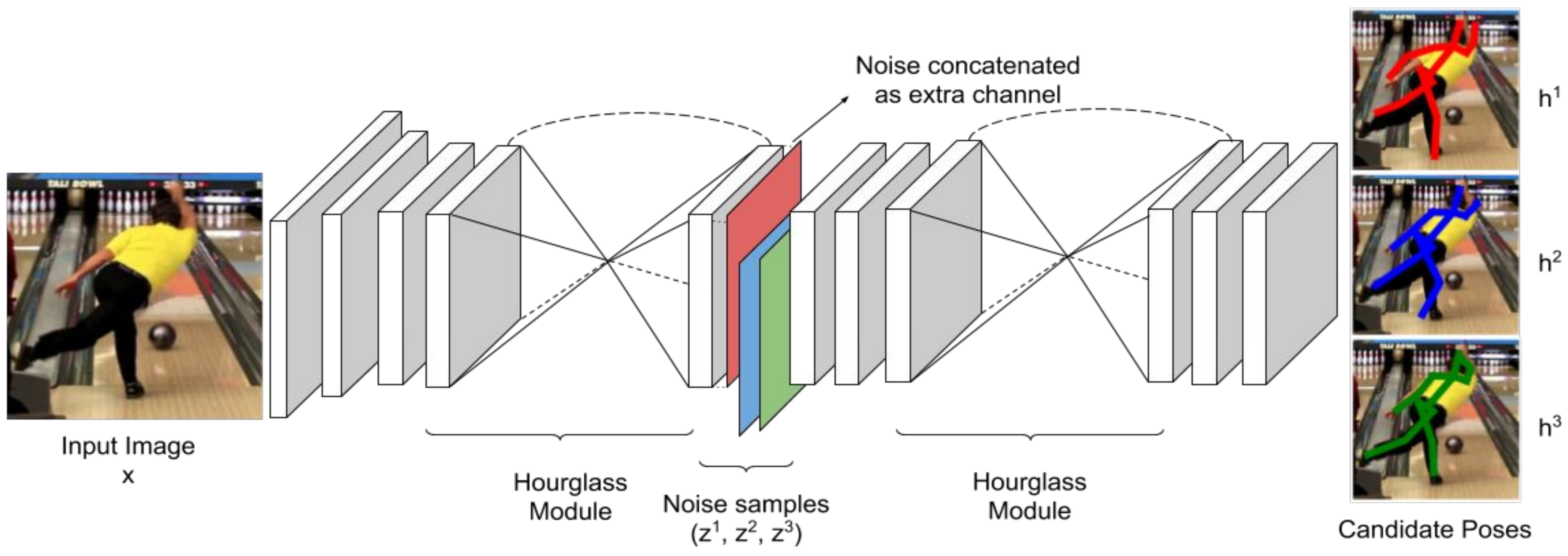


Fig: For a single input image **x** *and three different noise samples {*$\mathbf{z}^1, \mathbf{z}^2, \mathbf{z}^3$*} (represented as red, green and blue matrix respectively), the network produces three different candidate poses {*$\mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3$*}.*

Pointwise Prediction:

- Single input **x**.
- Multiple {$\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^k$} sampled from the probabilistic hourglass network.
- Optimal prediction according to the Δ with maximum expected utility (EU):

$$\mathbf{h}^*_\Delta(\mathbf{x};\mathbf{w}) = \arg\max_{k\in[1,K]} EU(\mathbf{h}^k) = \arg\min_{k\in[1,K]} \sum_{k'=1}^{K} \Delta(\mathbf{h}^k, \mathbf{h}^{k'})$$

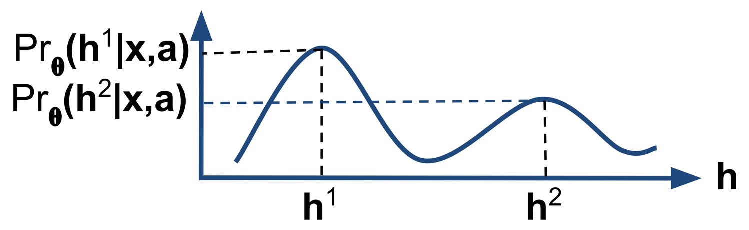
Learning Objective

- The **diversity coefficient** is the expected value of the loss.

$$\mathrm{DIV}_\Delta(\mathrm{Pr}_{\boldsymbol{\theta}}, \mathrm{Pr}_{\mathbf{w}}) = \sum_{\mathbf{h}^k, \mathbf{h}^{k'} \in \mathcal{H}} \Delta(\mathbf{h}^k, \mathbf{h}^{k'})\, \mathrm{Pr}_{\boldsymbol{\theta}}(\mathbf{h}^k)\, \mathrm{Pr}_{\mathbf{w}}(\mathbf{h}^{k'})$$

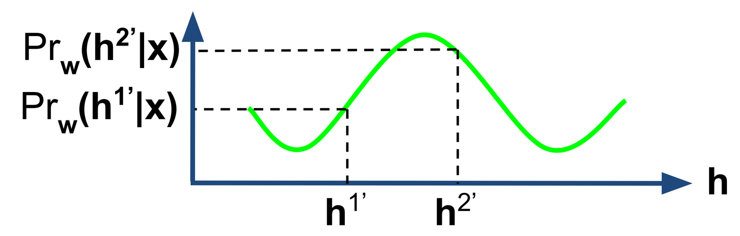
Diversity of a distribution $\mathrm{Pr}_{\boldsymbol{\theta}}$:

$$\mathrm{DIV}_\Delta(\mathrm{Pr}_{\boldsymbol{\theta}}, \mathrm{Pr}_{\mathbf{w}}) = \Delta(\mathbf{h}^1, \mathbf{h}^2)\, \mathrm{Pr}_{\boldsymbol{\theta}}(\mathbf{h}^1)\, \mathrm{Pr}_{\boldsymbol{\theta}}(\mathbf{h}^2)$$



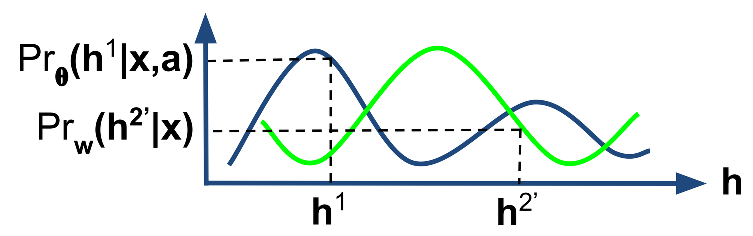
Diversity of a distribution $\mathrm{Pr}_{\mathbf{w}}$:

$$\mathrm{DIV}_\Delta(\mathrm{Pr}_{\mathbf{w}}, \mathrm{Pr}_{\mathbf{w}}) = \Delta(\mathbf{h}^{1'}, \mathbf{h}^{2'})\, \mathrm{Pr}_{\mathbf{w}}(\mathbf{h}^{1'})\, \mathrm{Pr}_{\mathbf{w}}(\mathbf{h}^{2'})$$



Diversity between the distributions $\mathrm{Pr}_{\boldsymbol{\theta}}$ and $\mathrm{Pr}_{\mathbf{w}}$:

$$\mathrm{DIV}_\Delta(\mathrm{Pr}_{\boldsymbol{\theta}}, \mathrm{Pr}_{\mathbf{w}}) = \Delta(\mathbf{h}^1, \mathbf{h}^{2'})\, \mathrm{Pr}_{\boldsymbol{\theta}}(\mathbf{h}^1)\, \mathrm{Pr}_{\mathbf{w}}(\mathbf{h}^{2'})$$



- We design a joint learning objective that minimizes the **dissimilarity coefficient**[4] (DISCO) between the prediction distribution and the conditional distribution.
- Formally, the dissimilarity coefficient is written as an affine combination of:
  - diversity between $\mathrm{Pr}_{\boldsymbol{\theta}}$ and $\mathrm{Pr}_{\mathbf{w}}$; and
  - diversity of each distribution.

$$\mathrm{DISC}_\Delta(\mathrm{Pr}_{\mathbf{w}}, \mathrm{Pr}_{\boldsymbol{\theta}}) = \mathrm{DIV}_\Delta(\mathrm{Pr}_{\mathbf{w}}, \mathrm{Pr}_{\boldsymbol{\theta}}) - \gamma \mathrm{DIV}_\Delta(\mathrm{Pr}_{\mathbf{w}}, \mathrm{Pr}_{\mathbf{w}}) - (1-\gamma)\mathrm{DIV}_\Delta(\mathrm{Pr}_{\boldsymbol{\theta}}, \mathrm{Pr}_{\boldsymbol{\theta}})$$

Optimization

We estimate the parameters of the two networks in three stages:

1. Supervised training of the two networks using small amount of ground truth pose data.
2. Iterative training: updating one network while keeping the other fixed on diverse data.
3. Jointly optimize both the networks together on diverse data.

Visualization of the effect of iterative learning of the two networks are shown below:

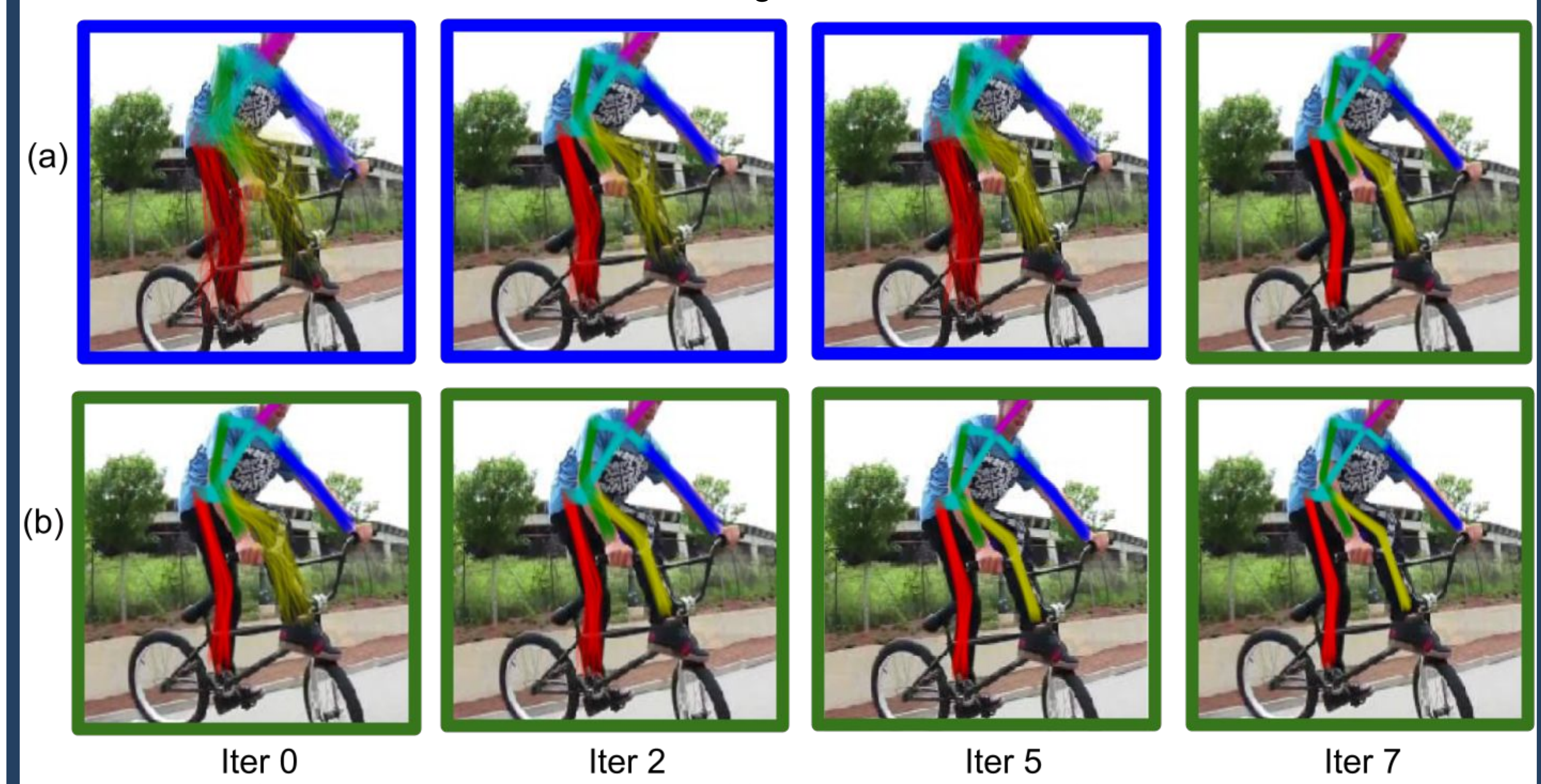
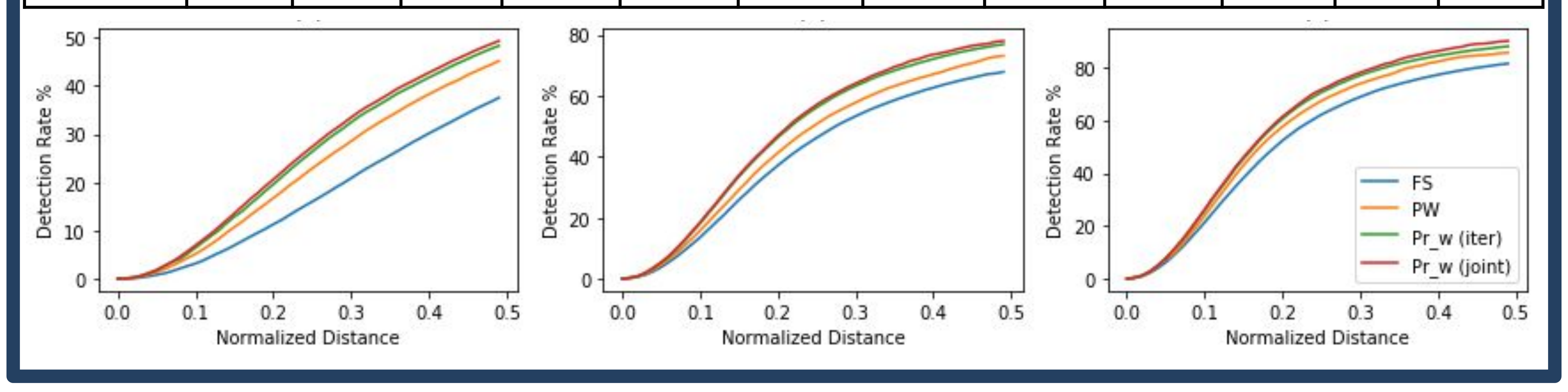


Fig: Example of superimposed pose predictions by (a) prediction network and (b) conditional network illustrating the uncertainty in the pose across training iterations.

Experiments and Results

- **Diverse data set:**
  - We use MPII Human Pose data set.
  - To obtain the various diverse data set, we choose three different data splits, {25−75,50−50,75−25}%, where we randomly discard 75%, 50%, and 25% of the pose annotations from the training images respectively.
- **Methods:**
  - Fully supervised (**FS**) stacked hourglass network trained on supervised subset.
  - Non probabilistic pointwise network (**PW**) trained with diverse data set.
  - Probabilistic DISCO-HG network (**Pr**) trained with diverse data set.
- **Results:**

Method	FS			PW			Pr_w (iter)			Pr_w (joint)		
Split	25%	50%	75%	25-75	50-50	75-25	25-75	50-50	75-25	25-75	50-50	75-25
PcKh @0.5	37.54	67.88	80.88	45.16	73.11	85.89	48.12	76.43	88.16	**49.41**	**78.01**	**90.21**



References

1. M. Pawan Kumar, Ben Packer,and Daphne Koller. *Modeling latent variable uncertainty for loss-based learning.* In ICML, 2012.
2. Diane Bouchacourt, M. Pawan Kumar,and Sebastian Nowozin. *DiscoNets: Dissimilarity coefficient networks.* In NIPS, 2016.
3. Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.
4. C. Radhakrishna Rao. *Diversity and dissimilarity coefficients: a unified approach.* Theoretical population biology, 21(1):24–43, 1982.