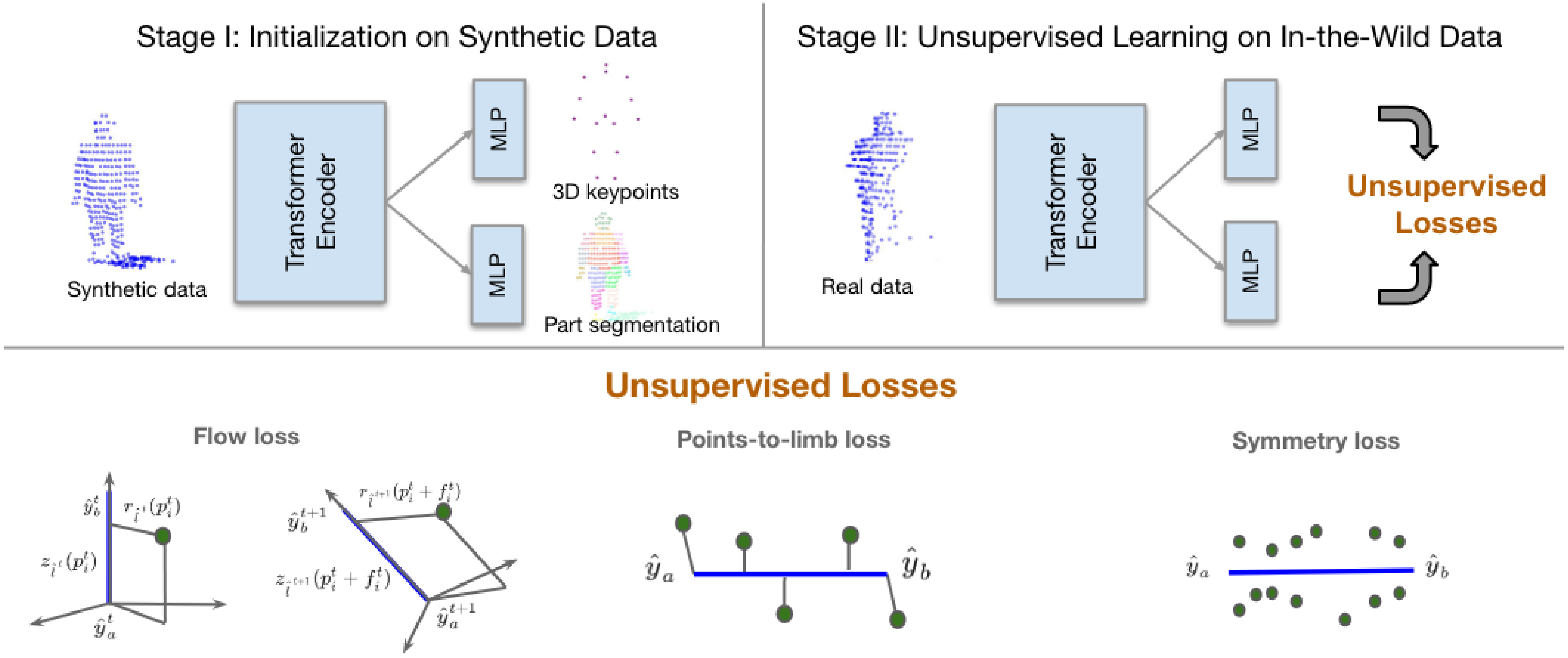


3D Human Keypoints Estimation from Point clouds in the Wild Without Human Labels

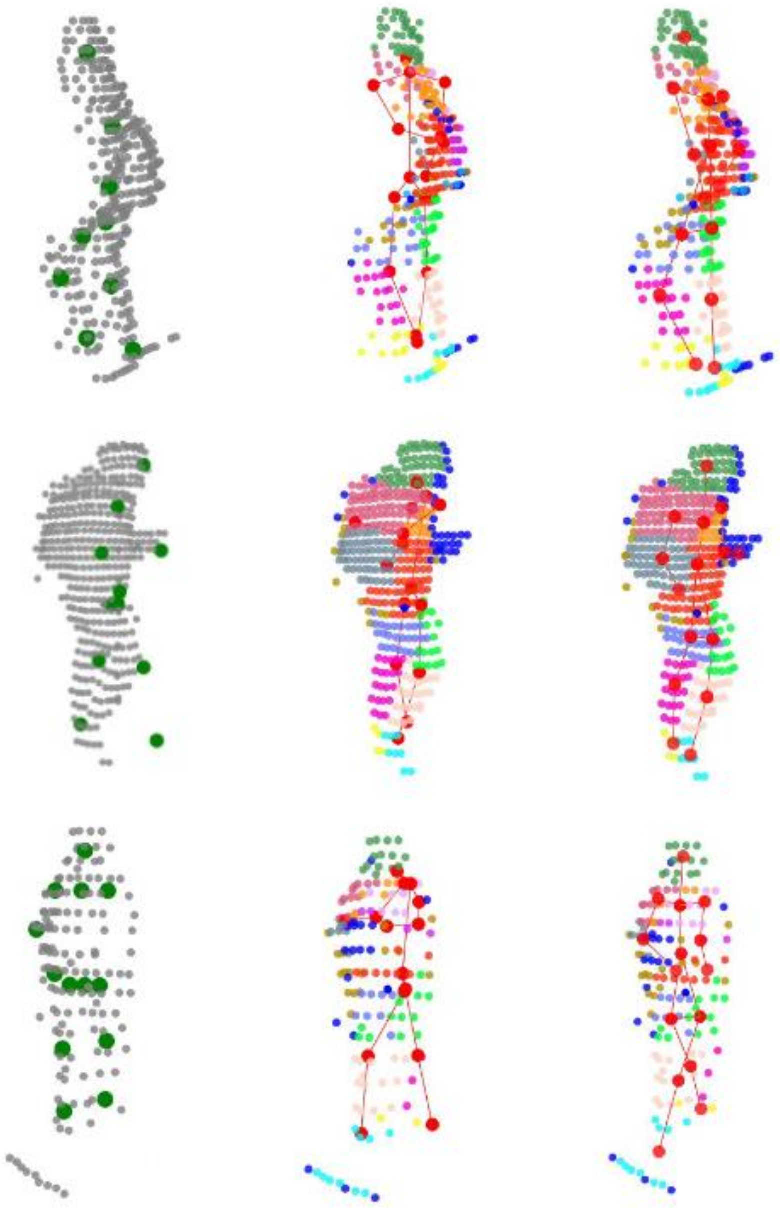
In CVPR 2023

Stanford University, Waymo

DaeYong Kim

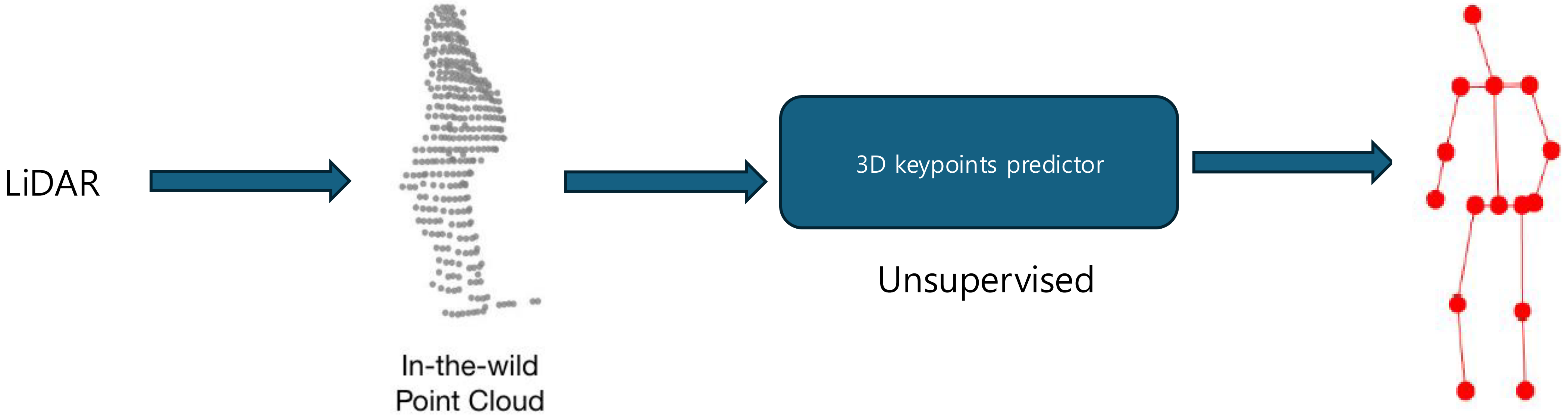
Dept. of Artificial Intelligence, Ajou University

Contents



Goal

3D human keypoints estimation from point clouds in the Wild

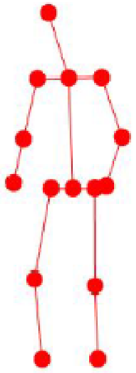


Prior works

Supervision

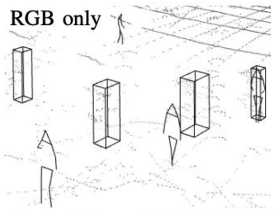
\+ Label

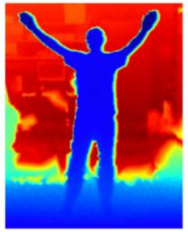
3D keypoints predictor



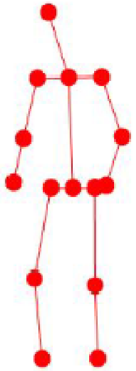
Prior work)
3d data + label : expensive
focus on utilizing 2d weak data

weak Supervision : RGB, depth, etc…





3D keypoints predictor



This work)
un-labeled LiDAR data has a lot of useful information

- *M. Furst, S. T. P. Gupta, R. Schuster, O. Wasenmuller, and D. Stricker, "HPERL: 3D human pose estimation from RGB and LiDAR," IEEE, 2021*
- *J. Zheng, X. Shi, A. Gorban, J. Mao, Y. Song, C. R. Qi, T. Liu, V. Chari, A. Cornman, Y. Zhou, and others, "Multi-modal 3D human pose estimation with 2D weak supervision in autonomous driving," IEEE, 2022*

Propose : GC-KPL

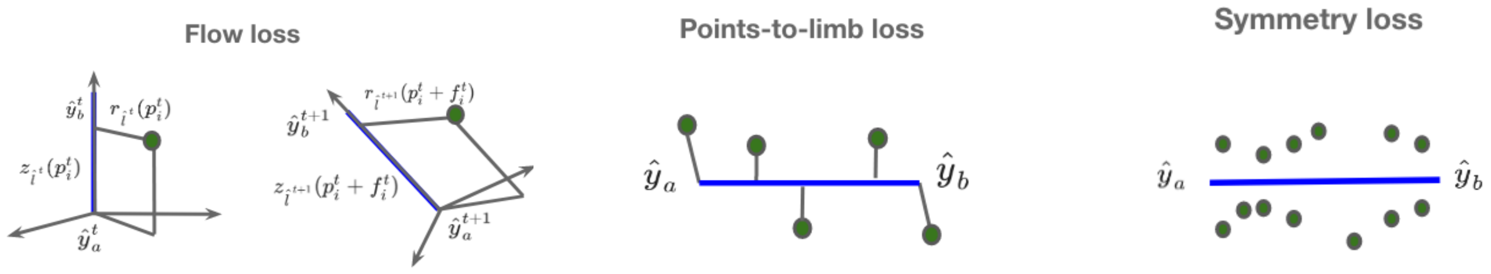
GC-KPL : Geometry Consistency inspired Key Point Leaning

Assumptions

- human skeletons are roughly centered rigid body parts
- surface points : “location + movement” -> skeleton “movement”

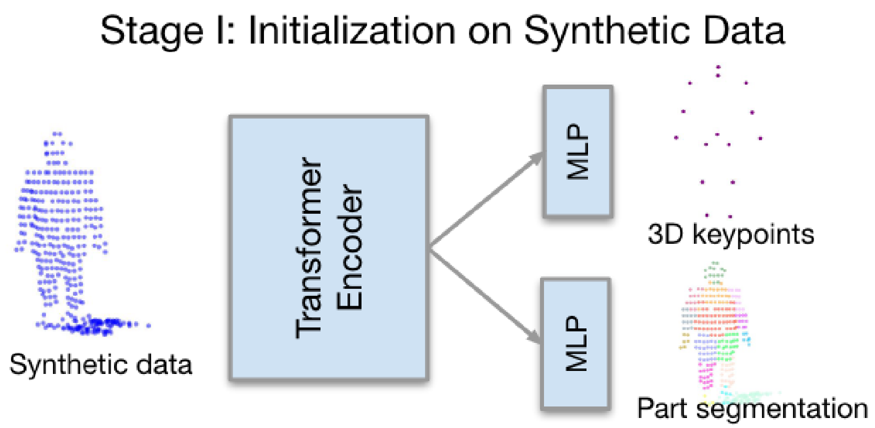
propose

- novel unsupervised losses term
  - Flow Loss
  - Points-to-Limb Loss
  - Symmetry Loss



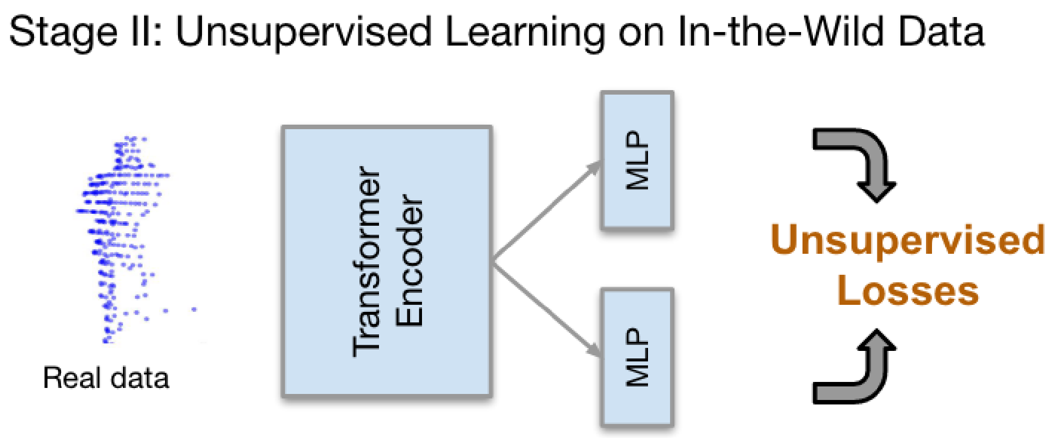
Process - Overall

Stage I – Warm Up



- transformer-based regression model : keypoints
- semantic segmentation model : localizing body parts on a synthetic data
- synthetic data : constructed from randomly posed **SMPL human body model**

Stage II – refine the network



- Using unsupervised Losses
- Flow Loss
- Points-to-limb Loss
- Symmetry Loss

Process – Stage I : Input & Output

Input : Point Clouds
Output : {3D keypoints, Part segmentation}

$$\mathbf{P} \in \mathbb{R}^{N\times 3} \rightarrow \hat{\mathbf{Y}} \in \mathbb{R}^{(J+1)\times 3}$$

$$\mathbf{P} \in \mathbb{R}^{N\times 3} \rightarrow \hat{\mathbf{W}} \in \mathbb{R}^{N\times (J+1)}$$

$$\{\hat{\mathbf{Y}}, \hat{\mathbf{W}}\} = f(\mathbf{P})$$

$$\forall i \in [N], \sum_{j=1}^{J+1} \hat{\mathbf{W}}_{i,j} = 1$$

$\hat{Y}$: 3D Locations of keyPoints
$\widehat{W}$: probability of each point i(body parts || background)

<Keypoints : L2 Loss>

$$\mathcal{L}_{kp} = \|\hat{\mathbf{Y}} - \mathbf{Y}\|_2$$

<segmentation : Cross-Entropy Loss>

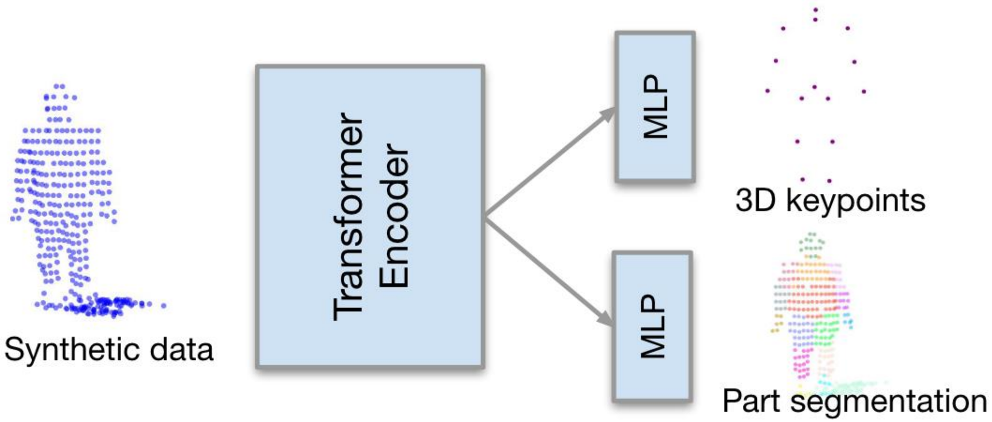
$$\mathcal{L}_{seg} = -\sum_{i=1}^{N}\sum_{j=1}^{J+1} \mathbf{W}_{i,j} \log(\hat{\mathbf{W}}_{i,j})$$

<Minimize>

$$\mathcal{L}_{syn} = \lambda_{kp}\mathcal{L}_{\mathrm{kp}} + \lambda_{seg}\mathcal{L}_{\mathrm{seg}}$$

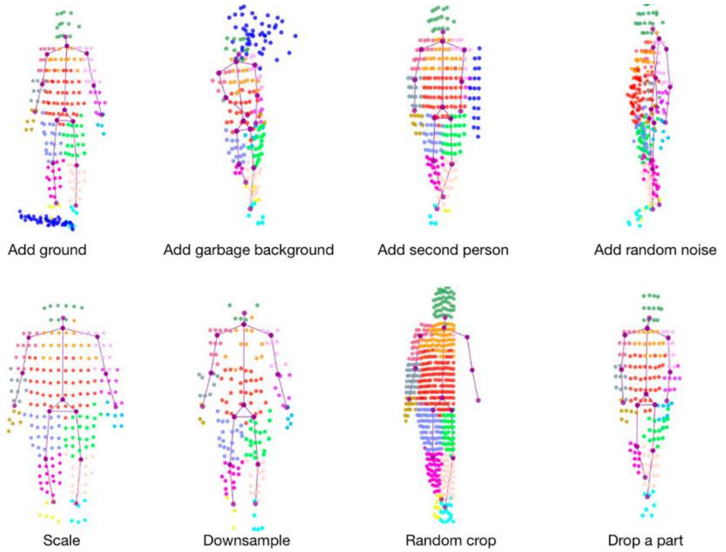
Process – Stage I

3D human keypoints estimation from points clouds in the wild



L2 -> regression

Cross Entropy -> classification

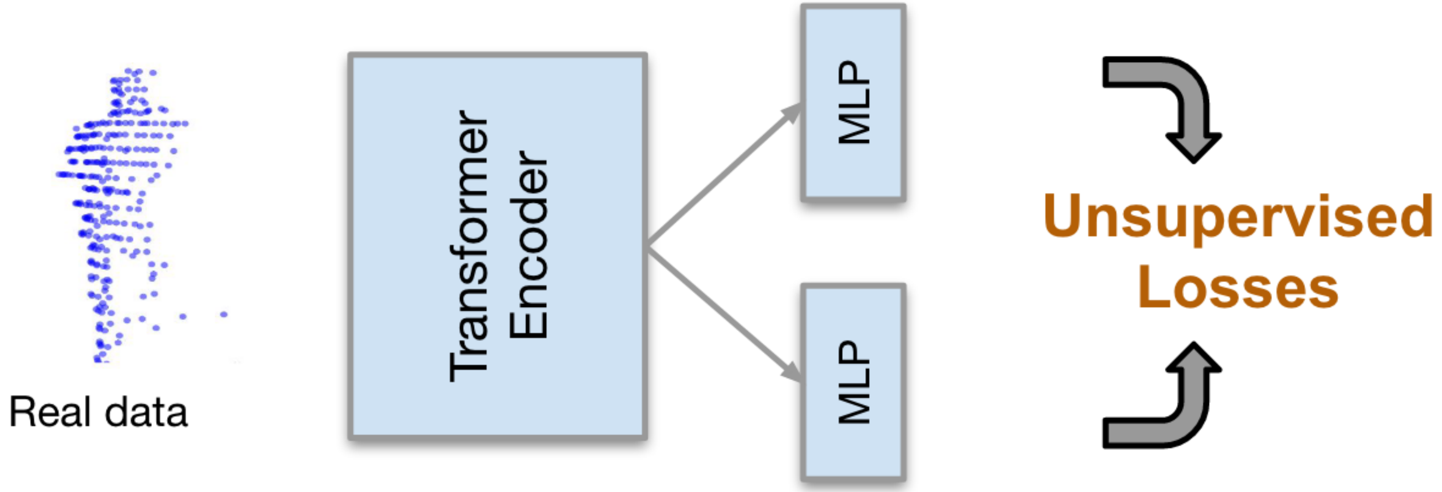


<augmentation>

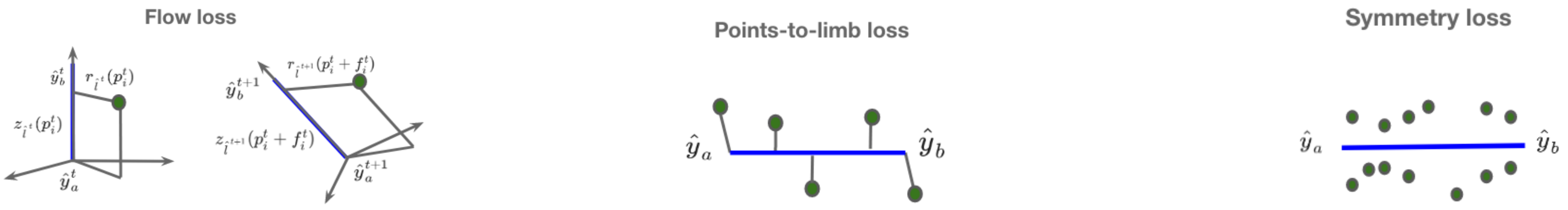
- Add ground
- Add garbage background
- Add random noise
- Scale
- Downsample
- Random crop
- Drop a part

Process – Stage II

3D human keypoints estimation from points clouds in the wild



points clouds -> set of points {x, y, z}



Losses

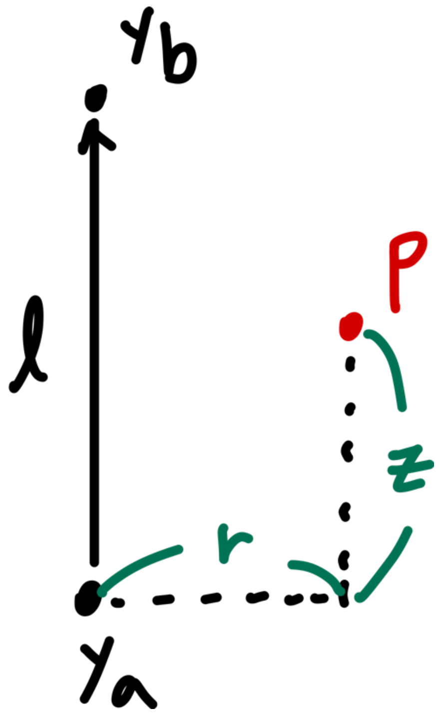
- L : human skeleton composed of limbs
- y_a : parent keypoint
- y_b : child keypoint

$$l = (y_a, y_b) \in L$$: Limb

$$\mathbf{z}(p, \hat{l}) = \frac{(p - \hat{y}_a) \cdot (\hat{y}_b - \hat{y}_a)}{||\hat{y}_b - \hat{y}_a||_2}$$: axial

$$\mathbf{r}(p, \hat{l}) = ||p - \hat{y}_a - \mathbf{z}(\hat{y}_b - \hat{y}_a, \hat{l})||_2$$: radial

$\hat{\mathbf{W}}_{ia}$: The probability of each point I belonging to body part a



points → each limbs' local cylindrical coordinate

Losses – Flow Loss

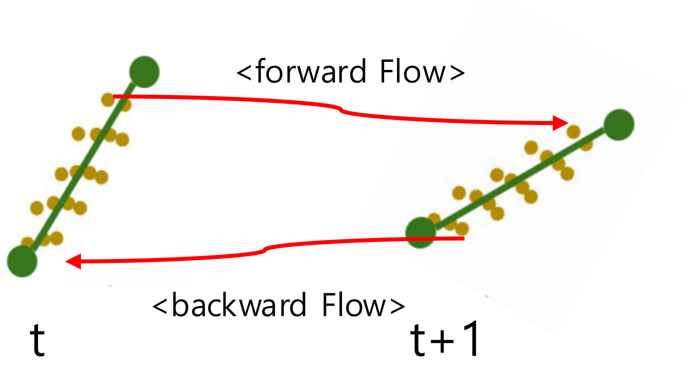
- For considering the predictions from two consecutive frames
- For consistency of the radial and altitude components of all points with respect to scene flow

<Forward Flow>

$$\mathcal{L}_{ff} = \frac{1}{N}\sum_i \hat{\mathbf{W}}_{ia}^t \cdot (|\mathbf{r}_{\hat{l}^{t+1}}(p_i^t + f_i^t) - \mathbf{r}_{\hat{l}^t}(p_i^t)| + |\mathbf{z}_{\hat{l}^{t+1}}(p_i^t + f_i^t) - \mathbf{z}_{\hat{l}^t}(p_i^t)|)$$

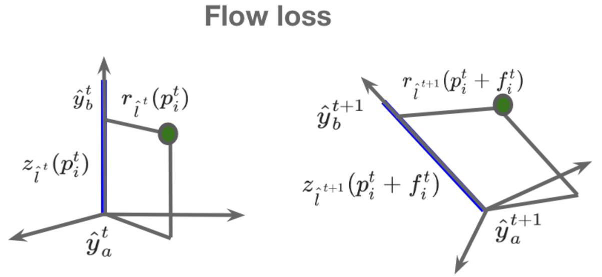
<backward Flow>

$$\mathcal{L}_{bf} = \frac{1}{N}\sum_i \hat{\mathbf{W}}_{ia}^{t+1} \cdot (|\mathbf{r}_{\hat{l}^t}(p_i^{t+1} + b_i^{t+1}) - \mathbf{r}_{\hat{l}^{t+1}}(p_i^{t+1})| + |\mathbf{z}_{\hat{l}^t}(p_i^{t+1} + b_i^{t+1}) - \mathbf{z}_{\hat{l}^{t+1}}(p_i^{t+1})|)$$



<Flow loss = Forward + backward>

$$\mathcal{L}_{flow} = \frac{1}{|L|}\sum_{\hat{l}^t} \frac{\mathcal{L}_{ff} + \mathcal{L}_{bf}}{2}$$



(a) After moving, **points** stay in the same place (despite rotation around axis) within each **limb's** local cylindrical coordinate system.

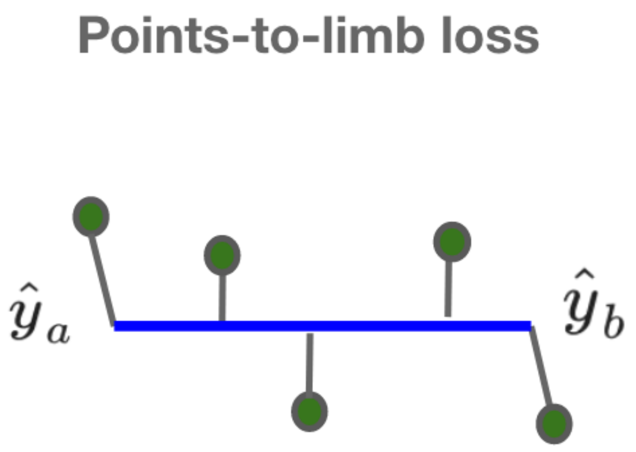
Losses – Points-to-Limb Loss

$$\mathcal{L}_{p2l}^{\hat{l}} = \frac{1}{N}\sum_{i}\hat{\mathbf{W}}_{ia}\mathbf{d}(p_i, \hat{l})$$

d : L2 distance function(point <-> limb)

<Points-to-Limb loss = Sum over all points>

$$\mathcal{L}_{\mathrm{p2l}} = \frac{1}{|L|}\sum_{\hat{l}}\mathcal{L}_{\mathrm{p2l}}^{\hat{l}}$$



(b) Minimize **points**-to-**limb** distance to encourage the limb to stay *within* the body.

- To minimize the distance of the points to the corresponding limb

Losses – Symmetry Loss

$$\mathcal{L}^{\hat{l}}_{sym} = \frac{1}{N}\sum_i \hat{\mathbf{W}}_{ia}(\mathbf{r}_{\hat{l}}(p_i) - \bar{\mathbf{r}}_{\hat{l}}(p_i))^2$$

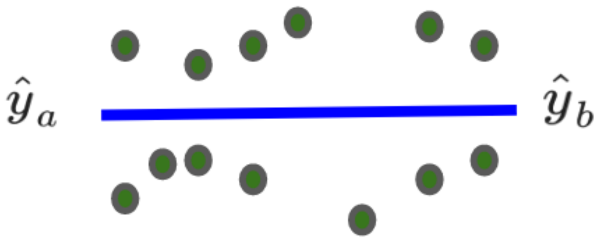
$\bar{\mathbf{r}}_{\hat{l}}$: weighted mean of radial values of points

$$\bar{\mathbf{r}}_{\hat{l}}(p_i) = \frac{\sum_j K_h(\mathbf{z}_{\hat{l}}(p_i), \mathbf{z}_{\hat{l}}(p_j))(\hat{\mathbf{W}}_{i*} \cdot \hat{\mathbf{W}}_{j*})\mathbf{r}_{\hat{l}}(p_j)}{\sum_j K_h(\mathbf{z}_{\hat{l}}(p_i), \mathbf{z}_{\hat{l}}(p_j))(\hat{\mathbf{W}}_{i*} \cdot \hat{\mathbf{W}}_{j*})}$$

K_h : Gaussian kernel with bandwith h

$$K_h(x,y) = e^{-(\frac{x-y}{h})^2}$$

Symmetry loss



(c) **Points** are symmetrical around **limb.** (i.e. points with similar height z have similar radius r)

<Symmetry loss = Sum over all points>

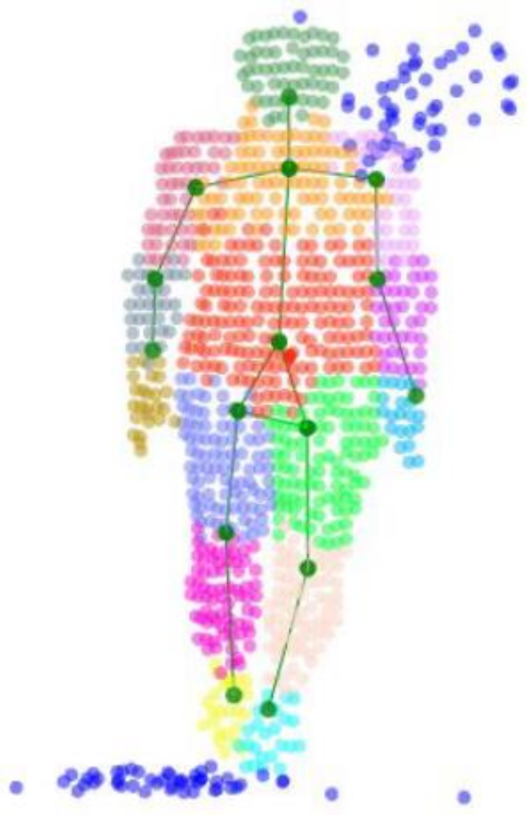
$$\mathcal{L}_{sym} = \frac{1}{|L|}\sum_{l \in L} \mathcal{L}^{l}_{sym}$$

- to encourage that all points around the limb are roughly symmetrical around it

Losses – Joint-to-Part Loss(Optional)

- To encourage each joint to be close to the center of the points on that part

$$\mathcal{L}_{j2p}^{j} = \left\| \hat{y}_j - \frac{\sum_i \hat{\mathbf{W}}_{ij} p_i}{\sum_i \hat{\mathbf{W}}_{ij}} \right\|_2$$

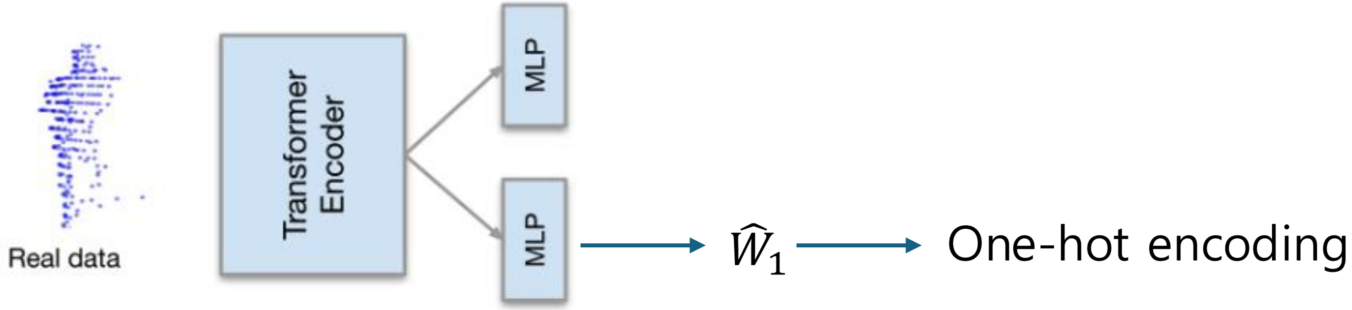


<Joint-to-part loss = Sum over all Joints>

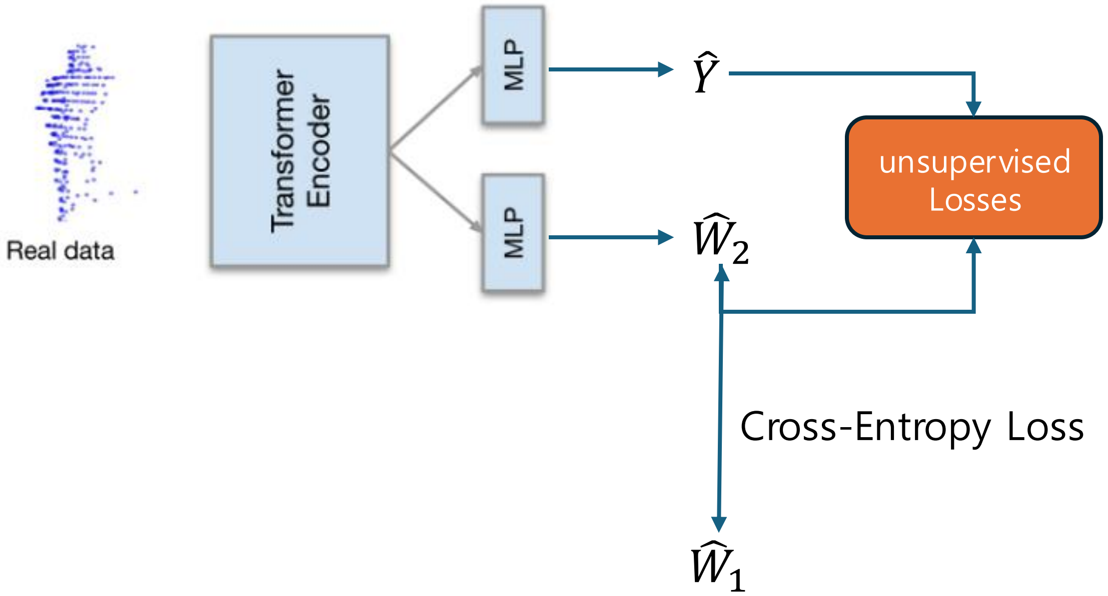
$$\mathcal{L}_{j2p} = \frac{1}{J} \sum_j \mathcal{L}_{j2p}^{j}$$

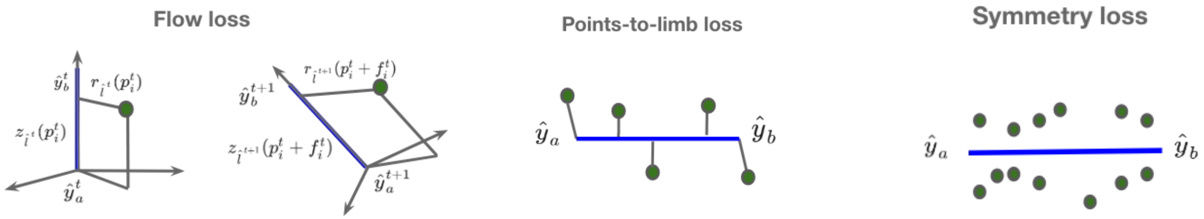
Losses – segmentation Loss && Training objective

Before Stage II



Stage II : Start Training

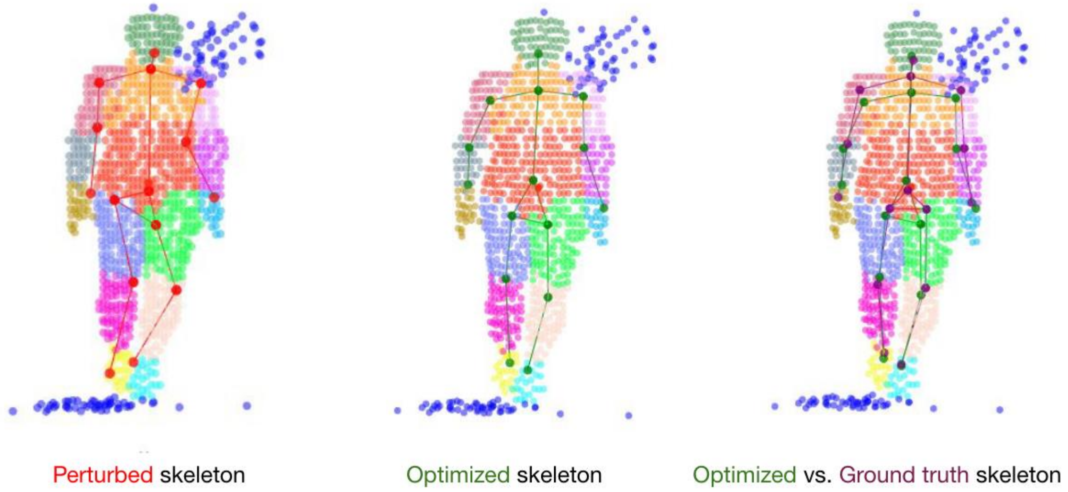


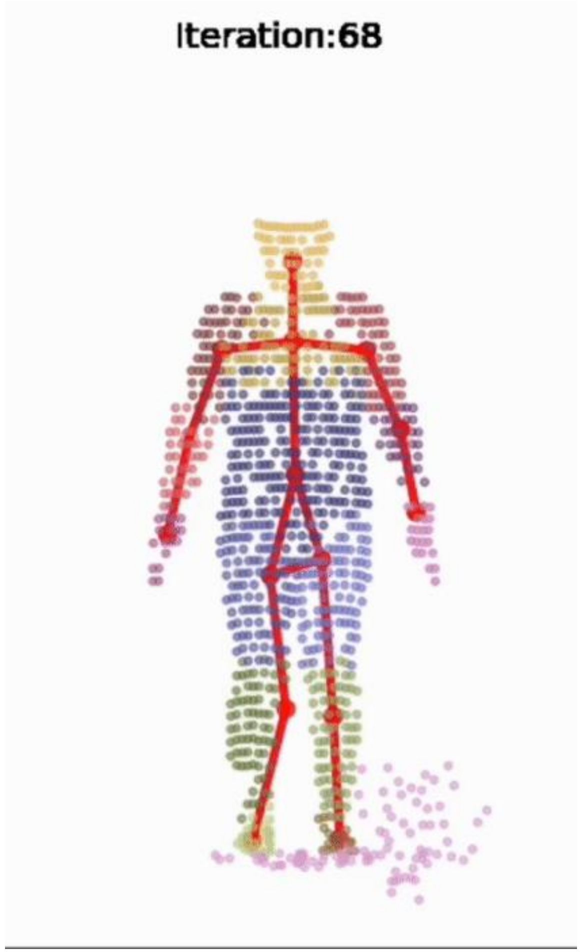


$$\mathcal{L} = \lambda_{flow}\mathcal{L}_{flow} + \lambda_{\text{p2l}}\mathcal{L}_{\text{p2l}} + \lambda_{sym}\mathcal{L}_{sym} + \lambda_{\text{j2p}}\mathcal{L}_{\text{j2p}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}$$

unsupervised Loss

Evaluation





hyper-parameter

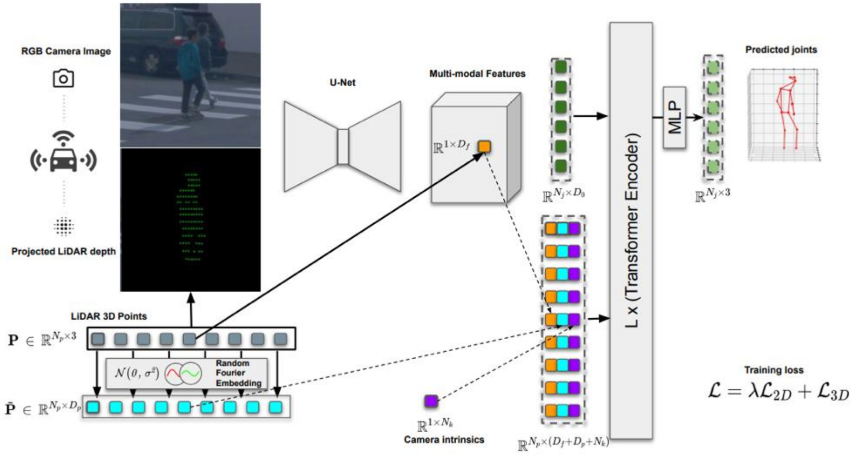
- $\lambda_{kp} = 0.5$
- $\lambda_{seg} = 1$
- $\lambda_{flow} = 0.02$
- $\lambda_{p2l} = 0.01$
- $\lambda_{sym} = 0.5$
- $\lambda_{j2p} = 2$
- $\lambda_{seg} = 0.5$

the result of applying the losses

$$\mathcal{L} = \lambda_{flow}\mathcal{L}_{flow} + \lambda_{\mathrm{p2l}}\mathcal{L}_{\mathrm{p2l}} + \lambda_{sym}\mathcal{L}_{sym} + \lambda_{\mathrm{j2p}}\mathcal{L}_{\mathrm{j2p}} + \lambda_{\mathrm{seg}}\mathcal{L}_{\mathrm{seg}}$$

Evaluation

Method	Backbone	Stage I supervised	1% training set MPJPE cm. (gain)	10% training set MPJPE cm. (gain)	100% training set MPJPE cm. (gain)
HUM3DIL [29]	Randomly initialized		19.57	16.36	12.21
GC-KPL	Pre-trained on synthetic only	✔	18.52 (-1.05)	15.10 (-1.26)	11.27 (-0.94)
	Pre-trained on 5,000 WOD-train	✔	17.87 (-1.70)	14.51 (-1.85)	10.73 (-1.48)
	Pre-trained on 200,000 WOD-train		17.80 (-1.77)	14.30 (-2.06)	10.60 (-1.61)
	Pre-trained on 200,000 WOD-train	✔	**17.20** (**-2.37**)	**13.40** (**-2.96**)	**10.10** (**-2.11**)



HUM3DIL : Image + LiDAR

Training data	$\text{MPJPE}_{\text{matched}}$ (↓)
Synthetic only	17.70
5,000 WOD-train	14.64
200,000 WOD-train	13.92

Table 2. Unsupervised learning (Stage II) results.

-A. Zanfir, M. Zanfir, A. Gorban, J. Ji, Y. Zhou, D. Anguelov, and C. Sminchisescu, "Hum3dil: Semi-supervised multi-modal 3D human pose estimation for autonomous driving" CoRL 2022