

VoxelKP: A Voxel-based Network Architecture for Human Keypoint Estimation in LiDAR Data

Supplementary Material

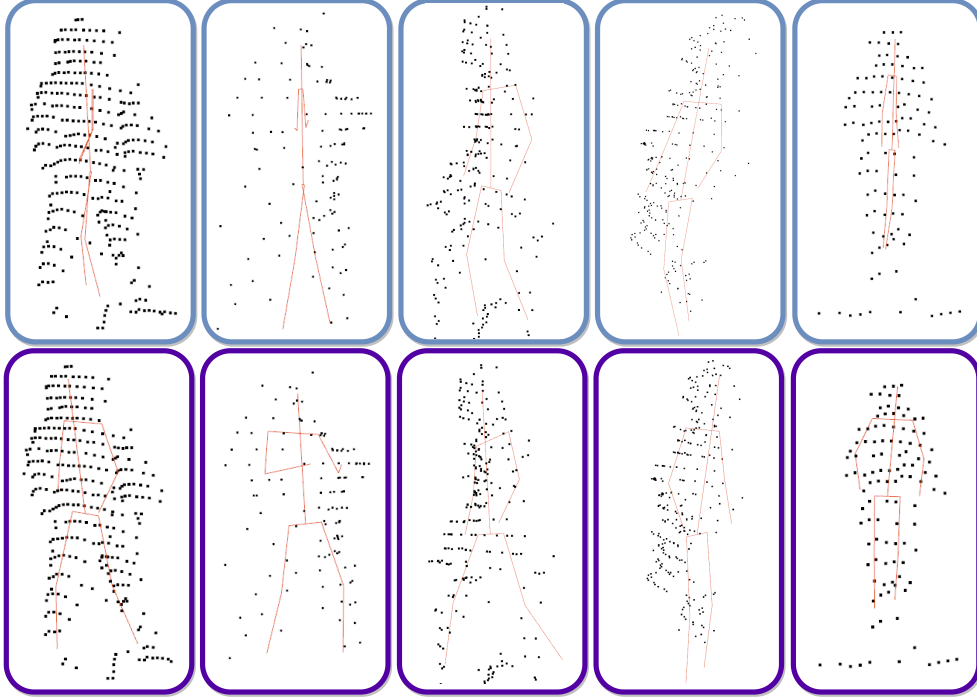


Figure 6: A visual demonstration of the baseline model (top row) and the proposed *VoxelKP* (bottom row) on matched human objects. Our method offers a more accurate estimation.

A TECHNICAL DETAILS

This section presents additional technical details of the network, loss functions, and metrics used.

A.1 SUPPLEMENTARY NETWORK DETAILS

The architecture of the stem module and prediction heads are presented in Figs. 7 and 8. The stem module includes CONV-BN-ReLU blocks with skip connections to extract low-level features. It contains one downsampling layer to obtain a smaller feature map. The model uses seven prediction heads. These heads predict: 1) the size of the bounding box, 2) the rotation of the bounding box, 3-5) the location of the box center and keypoints along the x, y, and z axes, 6) the visibility of keypoints, and 7) the Intersection over Union (IoU). Notably, we incorporate the IoU prediction to enhance performance, following Hu et al. (2022).

A.2 LOSSES

We use three types of losses in our work including the skeleton loss. Notably, the ground truth annotations are converted into the same sparse representation as the predictions for loss computation.

Heatmap Loss Our network outputs a set of heatmaps, one per class. This heatmap encoding allows our model to classify and localize objects in 3D space simultaneously. In the training phase, we assign positive heatmap indices based on ground truth annotations. Specifically, we identify the voxel closest to the annotated bounding box center and mark that voxel with a positive heatmap

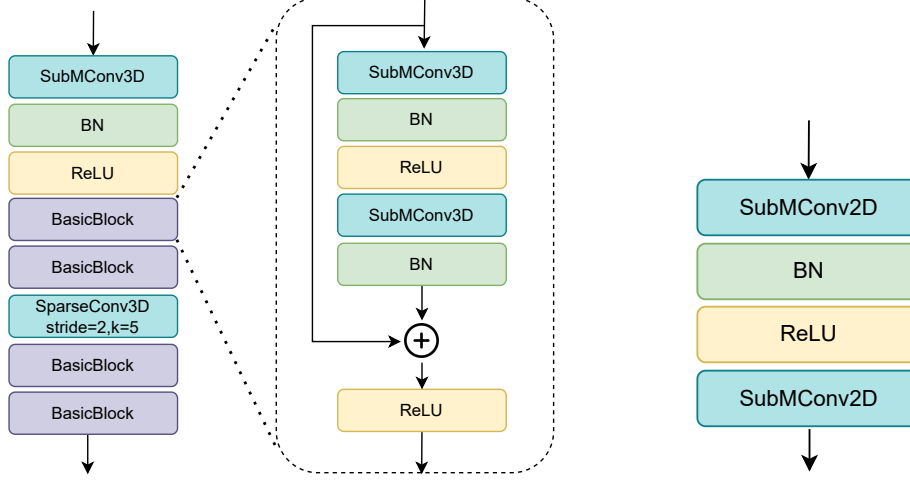


Figure 7: The architecture of the stem module. Figure 8: The architecture of prediction heads.

value. We supervise these heatmaps using an adapted focal loss function Lin et al. (2017); Law & Deng (2018); Chen et al. (2023). With the annotated and predicted heatmaps I and \hat{I} , we have:

$$FL(I, \hat{I}) = \frac{-1}{N} \sum_{c=1}^C \sum_{v=1}^V \begin{cases} (1 - \hat{I})^\alpha \cdot \log(\hat{I}), & \text{if } I = 1 \\ \log(1 - \hat{I}) \cdot \hat{I}^\alpha \cdot (1 - I)^\beta, & \text{otherwise} \end{cases}, \quad (3)$$

where N , C , V are the batch size, number of channels, and number of voxels, respectively. α and β are the hyper-parameters to weigh each voxel. We use $\alpha = 2$ and $\beta = 4$ in this work, following Law & Deng (2018).

L1 Regression Loss We adopt a simple L1 loss for other prediction heads of coordinates and key-point visibilities. With the ground truth and predicted values Y and \hat{Y} , we have:

$$L1(Y, \hat{Y}) = \frac{1}{N} \sum_{c=1}^C \|Y - \hat{Y}\|_1. \quad (4)$$

Skeleton Regularization We propose to use a skeleton loss to encode prior information about the relative positioning of keypoints. For this purpose, we include bone length regularization in the loss function. This term computes the distance between the ground truth bone length and the predicted bone length. Specifically, given the ground truth keypoint locations Y and predicted keypoint locations \hat{Y} , we first compute the skeleton bone lengths $BL(Y)$ and $BL(\hat{Y})$ by calculating the Euclidean distance between connected keypoint pairs. The skeleton loss is then calculated as the Huber loss $h(\cdot)$ between the predicted bone lengths $BL(\hat{Y})$ and ground truth bone lengths $BL(Y)$, resulting in:

$$SK(\hat{Y}, Y) = h(BL(\hat{Y}), BL(Y)). \quad (5)$$

This enforces the model to predict keypoint locations that respect the biomechanical constraints of bone lengths in the human skeleton. Matching the distribution of predicted bone lengths to the ground truth, ensures awareness of the spatial relationships between different joints. The skeleton loss penalizes predicted keypoints that violate the physical constraints of bone lengths, acting as a strong prior for plausible human poses.

A.3 METRICS

We use mean per-joint position error (MPJPE), pose estimation metric (PEM), and object keypoint similarity (OKS) to evaluate our method. Formally, let $\hat{Y} \in \mathbb{R}^{J \times 3}$ be the predicted keypoints of a human, $Y \in \mathbb{R}^{J \times 3}$ be the ground truth, and $v_j \in 0, 1$ be the visibility of each joint j . The MPJPE

metric is defined as:

$$\text{MPJPE}(Y, \hat{Y}) = \frac{1}{\sum_j v_j} \sum_{j \in [J]} v_j \|y_j - \hat{y}_j\|_2. \quad (6)$$

Note that MPJPE requires a one-to-one match between the keypoints predictions and ground truth. Therefore, a Hungarian matching is performed to match the predicted and annotated keypoints before calculating the MPJPE.

PEM further takes into account the matching accuracy that is essentially a sum of the MPJPE over visible matched keypoints with a penalty for unmatched keypoints. Note that the unmatched keypoints include both the ground truth keypoints without matching predicted keypoints and the predicted keypoints without matching ground truth objects.

$$\text{PEM}(Y, \hat{Y}) = \frac{\sum_{i \in M} \|y_i - \hat{y}_i\|_2 + C|U|}{|M| + |U|}, \quad (7)$$

where M is a set of indices of matched keypoints, $|U|$ is a set of indices of unmatched keypoints, and $C = 0.25$ is a constant penalty for an unmatched keypoint.

Additionally, we include the classic metric of OKS in this work. The OKS metric is not computed per keypoint, it is a relative metric computed for each human body. In OKS, each ground truth object also has a scale s which we define as the square root of the object segment area. OKS is computed as the arithmetic average across all labeled keypoints in an instance.

$$\text{OKS} = \frac{\sum_j e^{-\frac{d_j^2}{2s^2k_j^2}v_j}}{\sum_i v_i} \quad (8)$$

where d_j is the Euclidean distance between each corresponding ground truth and detected keypoint, k_j is a per-joint constant provided by COCO Lin et al. (2014). The reported OKS@KP is averaged over multiple OKS values, which are calculated for OKS thresholds starting at 0.50, increasing in steps of 0.05, and ending at 0.95.

B ADDITIONAL RESULTS

B.1 FULL EVALUATION

We report the full spectrum of the evaluation, including MPJPE, OKS@AP, and PEM. The details for each metric can be found in Appendix A.3.

part	Head	Shoulders	Elbows	Wrists	Hips	Knees	Ankles	All
MPJPE	0.0570	0.0669	0.0948	0.1467	0.0670	0.0820	0.1084	0.0887
OKS@AP	0.6393	0.8917	0.7197	0.3791	0.9533	0.8586	0.7581	0.7300
PEM	0.1569	0.1563	0.1746	0.1987	0.1576	0.1660	0.1765	0.1695

Table 6: Full evaluation of *VoxelKP*.