PoseKernelLifter: Metric Lifting of 3D Human Pose using Sound

Zhijian Yang1,2*, Xiaoran Fan1, Volkan Isler1,3*, Hyun Soo Park1,3*

1Samsung AI Center NY, 2University of Illinois Urbana Champaign, 3University of Minnesota Twin Cities

Figure 1. We present a new method for metric reconstruction of a human’s pose from a single image along with audio signals transmitted from consumer-grade speakers. Our method recovers the metric scale by leveraging the fact that sound travels at a constant speed through a fixed medium.

Abstract

Reconstructing the 3D pose of a person in metric scale from a single view image is a geometrically ill-posed problem. For example, we cannot measure the exact distance of a person to the camera from a single view image without additional scene assumptions (e.g., known height). Existing learning based approaches circumvent this issue by reconstructing the 3D pose up to scale. However, there are many applications such as virtual telepresence, robotics, and augmented reality that require metric scale reconstruction. In this paper, we show that audio signals recorded along with an image, provide complementary information to reconstruct the metric 3D pose of the person. The key insight is that as the audio signals traverse across the 3D space, their interactions with the body provide metric information about the body’s pose. Based on this insight, we introduce a time-invariant transfer function called pose kernel—the impulse response of audio signals induced by the body pose. The main properties of the pose kernel are that (1) its envelope highly correlates with 3D pose, (2) the time response corresponds to arrival time, indicating the metric distance to the microphone, and (3) it is invariant to changes in the scene geometry configurations. Therefore, it is readily generalizable to unseen scenes. We design a multi-stage 3D CNN that fuses audio and visual signals and learns to reconstruct 3D pose in a metric scale. We show that our multi-modal method produces accurate metric reconstruction in real-world scenes, which is not possible with state-of-the-art lifting approaches including parametric mesh regression and depth regression.

1. Introduction

Since the projection of the 3D world onto an image loses scale information, 3D reconstruction of a human’s pose from a single image is an ill-posed problem. To address this limitation, human pose priors have been used in existing lifting approaches [6, 20, 31, 34, 43, 60, 75] to reconstruct the plausible 3D pose given the 2D detected pose by predicting relative depths. The resulting reconstruction, nonetheless, still lacks metric scale, i.e., the metric scale cannot be recovered without making an additional assumption such as known height or ground plane contact. This fundamental limitation of 3D pose lifting precludes applying it to real-world downstream tasks, e.g., smart home facilitation, robotics, and augmented reality, where the precise metric measurements of human activities are critical, in relation to the surrounding physical objects.

In this paper, we study a problem of metric human pose reconstruction from a single view image by incorporating a new sensing modality—audio signals from consumer-grade speakers (Figure 1). Our insight is that while traversing a 3D environment, the transmitted audio signals undergo a
characteristic transformation induced by the geometry of reflective physical objects including human body. This transformation is subtle yet highly indicative of body pose geometry, which can be used to reason about the metric scale reconstruction. For instance, the same music playing in a room sounds differently based on the presence or absence of a person, and more importantly, as the person moves.

We parametrize this transformation of audio signals using a time-invariant transfer function called pose kernel—an impulse response of audio induced by a body pose, i.e., the received audio signal is a temporal convolution of the transmitted signal with the pose kernel. Three key properties of pose kernel enables metric 3D pose lifting in a generalizable fashion: (1) metric property: its impulse response is equivalent to the arrival time of the reflected audio, and therefore, it provides metric distance from the receiver (microphone); (2) uniqueness: the envelope of pose kernel is strongly correlated with the location and pose of the target person; (3) invariance: it is invariant to the geometry of surrounding environments, which allows us to generalize it to unseen environments.

While highly indicative of pose and location of the person in 3D, the pose kernel is a time-domain signal. Integrating it with spatial-domain 2D pose detection is non-trivial. Further, generalization to new scenes requires precise 3D reasoning where existing audio-visual learning tasks such as source separation in an image domain and image representation learning [13, 16, 42, 59] are not applicable.

We address this challenge in 3D reasoning of visual and audio signals, by learning to fuse the pose kernels from multiple microphones and the 2D pose detected from an image, using a 3D convolutional neural network (3D CNN): (1) we project each point in 3D onto the image to encode the likelihood of landmarks (visual features); and (2) we spatially encode the time-domain pose kernel in 3D to form audio features. Inspired by the convolutional pose machine architecture [70], a multi-stage 3D CNN is designed to predict the 3D heatmaps of the joints given the visual and audio features. This multi-stage design increases effective receptive field with a small convolutional kernel (e.g., features). This multi-stage design increases effective receptive fields is used to

Why Metric Scale? Smart home technology is poised to enter our daily activities, in particular, for monitoring fragile populations including children, patients, and the elderly. This requires not only 3D pose reconstruction but also holistic 3D understanding in the context of metric scenes, which allows AI and autonomous agents to respond in a situation-aware manner. While multiview cameras can provide metric reconstruction, the number of required cameras to cover the space increases quadratically as area increases. Our novel multi-modal solution can mitigate this challenge by leveraging multi-source audios (often inaudible) generated by consumer grade speakers (e.g., Alexa).

Contributions This paper makes a major conceptual contribution that sheds a new light on a single view pose estimation by incorporating with audio signals. The technical contributions include (1) a new formulation of the pose kernel that is a function of the body pose and location, which can be generalized to a new scene geometry, (2) the spatial encoding of pose kernel that facilitates fusing visual and audio features, (3) a multi-stage 3D CNN architecture that can effectively fuse them together, and (4) a strong performance of our method, outperforming state-of-the-art lifting approaches with meaningful margin.

2. Related work

This paper is primarily concerned with integrating information from audio signals with single view 3D pose estimation to obtain metric scale. We briefly review the related work in these domains.

Vision based Lifting While reconstructing 3D pose (a set of body landmarks) from a 2D image is geometrically ill-posed, the spatial relationship between landmarks provides a geometric cue to reconstruct the 3D pose [57]. This relationship can be learned from datasets that include 2D and 3D correspondences such as Human3.6M [22], MPI-INF-3DHP [36] (multiview), Surreal [62] (synthetic), and 3DPW [65] (external sensors). Given the 3D supervision, the spatial relationship can be directly learned via supervised learning [6, 20, 56, 60]. Various representations have been proposed to effectively encode the spatial relationship such as volumetric representation [43], graph structure [4, 11, 72, 77], transformer architecture [31, 34, 75], compact designs for realtime reconstruction [37, 38], and inverse kinematics [30]. These supervised learning approaches that rely on the 3D ground truth supervision, however, show limited generalization to images of out-of-distribution scenes and poses due to the domain gap. Weakly supervised, self-supervised, and unsupervised learning have been used to address this challenge. For instance, human poses in videos are expected to move and deform continuously over time, leading to a temporal self-supervision [45]. A dilated convolution that increases temporal receptive fields is used to
learn the temporal smoothness \cite{44, 61}, a global optimization is used to reconstruct temporally coherent pose and camera poses \cite{2}, and spatio-temporal graph convolution is used to capture pose and time dependency \cite{4, 9, 33}. Multiview images provide a geometric constraint that allows learning view-invariant visual features to reconstruct 3D pose. The predicted 3D pose can be projected onto other view images \cite{46, 47, 65}, stereo images are used to triangulate a 3D pose which can be used for 3D pseudo ground truth for other views \cite{23, 24, 28}, and epipolar geometry is used to learn 2D view invariant features for reconstruction \cite{21, 74}. Adversarial learning enables decoupling the 3D poses and 2D images, i.e., 3D reconstruction from a 2D image must follow the distribution of 3D poses, which allows learning from diverse images (not necessarily videos or multiview) \cite{8, 29, 64}. A characterization and differentiable augmentation of datasets, further, improves the generalization \cite{18, 68}. With a few exceptions, despite remarkable performance, the reconstructed poses lack the metric scale because of the fundamental ambiguity of 3D pose estimation. Our approach leverages sound generated by consumer-grade speakers to lift the pose in 3D with physical scale.

**Multimodal Reconstruction** Different modalities have been exploited for the purpose of 3D sensing and reconstruction, include RF based \cite{19, 25, 26, 78, 79}, inertial based \cite{54, 73}, and acoustic based \cite{7, 10, 14, 15, 53, 71, 76}. Various applications including self-driving car \cite{19}, robot manipulation and grasping \cite{40, 66, 67, 69}, simultaneous localization and mapping (SLAM) \cite{1, 12, 52, 55, 58} benefited from multimodal reconstruction. Audio, given its ambient nature, has attracted unique attention in multimodal machine learning \cite{3, 17, 32, 39, 41, 48}. However, few works \cite{10, 71, 76} appear in the area of multimodal geometry understanding using audio as a modality, due to the heavy audio multipath posing various difficulties in 3D understanding. Human pose, given its diverse nature, is especially challenging for traditional acoustic sensing, thus is sparsely studied. While similar signals like WiFi and FMCW radio have been used for human pose estimation \cite{25, 78, 79}, audio signal, given its lower speed of propagation, offers more accurate distance measurement than RF-based.

We address the challenge of audio multipath and uncover the potential of audio in accurate metric scale 3D human pose estimation. Specifically, we present the first method that combines audio signals with the 2D pose detection to reason about the 3D spatial relationship for metric reconstruction. Our approach is likely to be beneficial for various applications including smart home, AR/VR, robotics.

### 3. Method

We make use of audio signals as a new modality for metric human pose estimation. We learn a pose kernel that transforms audio signals, which can be encoded in 3D in conjunction with visual pose prediction as shown in Figure 2.

#### 3.1. Pose Kernel Lifting

We cast the problem of 3D pose lifting as learning a function \(g_\theta\) that predicts a set of 3D heatmaps \(\{P_i\}_{i=1}^N\) given an input image \(I \in [0, 1]^{W \times H \times 3}\) where \(P_i : \mathbb{R}^3 \rightarrow [0, 1]\) is the likelihood of the \(i^{th}\) landmark over a 3D space, \(W\) and \(H\) are the width and height of the image, respectively, and \(N\) is the number of landmarks. In other words,

\[
\{P_i\}_{i=1}^N = g_\theta(I),
\]

where \(g_\theta\) a learnable function parametrized by its weights \(\theta\) that lift a 2D image to the 3D pose. Given the predicted 3D heatmaps, the optimal 3D pose is given by \(X_1^* = \...\)
argmax $P_i(X)$ so that $X^*_i$ is the optimal location of the $i^{th}$ landmark. In practice, we use a regular voxel grid to represent $P$.

We extend Equation (1) by leveraging audio signals to reconstruct a metric scale human pose, i.e.,

$$\{P_j\}_{j=1}^N = g_0(I, \{k_j(t)\}_{j=1}^M),$$

where $k_j(t)$ is the pose kernel heard from the $j^{th}$ microphone—a time-invariant audio impulse response with respect to human pose geometry that transforms the transmitted audio signals, as shown in Figure 2. $M$ denotes the number of received audio signals'. The pose kernel transforms the transmitted waveform as follows:

$$r_j(t) = s(t) * (K_j(t) + k_j(t)), \tag{3}$$

where * is the operation of time convolution, $s(t)$ is the transmitted source signal and $r_j(t)$ is the received signal at the location of the $j^{th}$ microphone. $K_j(t)$ is the empty room impulse response that accounts for transformation of the source signal due to the static scene geometry, e.g., wall and objects, in the absence of a person. $k_j(t)$ is the pose kernel measured at the $j^{th}$ microphone location that accounts for signal transformation due to human pose.

The pose kernel can be obtained using the inverse Fourier transform, i.e.,

$$k_j(t) = F^{-1}\{K_j(f)\}, \quad K_j(f) = \frac{R_j(f)}{S(f)} - \overline{K}_j(f), \tag{4}$$

where $F^{-1}$ is the inverse Fourier transformation, and $R_j(f)$, $S(f)$, and $\overline{K}_j(f)$ are the frequency responses of $r_j(t)$, $s(t)$, and $K_j(t)$, respectively, e.g., $R(f) = F[r(t)]$.

Since the pose kernel is dominated by direct reflection from the body, it is agnostic to scene geometry\(^1\). The scene geometry is factored out by the empty room impulse response $\overline{K}_j(t)$ and the source audios $s(t)$ are canceled by the received audios $r(t)$, which allows us to generalize the learned $g_0$ to various scenes.

### 3.2. Spatial Encoding of Pose Kernel

We encode the time-domain pose kernel of the $j^{th}$ microphone, $k_j(t)$ to 3D spatial-domain where audio and visual signals can be fused. A transmitted audio at the speaker’s location $s_{spk} \in \mathbb{R}^3$ reverts back by the body surface at $X \in \mathbb{R}^3$ and arrives at the microphone’s location $s_{mic} \in \mathbb{R}^3$. The arrival time is:

$$t_X = \frac{\|s_{spk} - X\| + \|s_{mic} - X\|}{v}, \tag{5}$$

where $t$ is the arrival time, and $v$ is the constant speed of sound (Figure 3).

The pose kernel is a superposition of impulse responses from the reflective points in the body surface, i.e.,

$$k_j(t) = \sum_{X \in \mathcal{X}} A(X) \delta(t - t_X), \tag{6}$$

where $\delta(t - t_X)$ is the Dirac delta function (impulse response) at $t = t_X$. $t_X$ is the arrival time of the audio signal reflected by the point $X$ on the body surface $\mathcal{X}$. $A(X)$ is the reflection coefficient (gain) at $X$.

Equation (5) and (6) imply two important spatial properties of the pose kernel. (i) Since the locus of points whose sum of distances to the microphone and the speaker is an ellipsoid, Equation (5) implies that the same impulse response can be generated by any point on this ellipsoid. (ii) Due to the constant speed of sound, the response of the arrival time can be interpreted as that of the spatial distance by evaluating the pose kernel at the corresponding arrival time, $t_X$:

$$K_j(X) = k_j(t)|_{t=t_X}.$$  \tag{7}

where $K_j(X)$ is the spatial encoding of the pose kernel at $X \in \mathbb{R}^3$.

Let us illustrate the spatial encoding of pose kernel. Consider a point object $X \in \mathbb{R}^3$ that reflects an audio signal from the speaker $s_{spk}$ which is received by the microphone $s_{mic}$ as shown in Figure 3. The received audio is delayed by $t_X$, which can be represented as a pose kernel $k(t) = A(X) \delta(t - t_X)$. This pose kernel can be spatially encoded as $K(X)$ because the speed of the sound is constant. Note that there exists the infinite number of possible locations of $X$ given the pose kernel because any point (e.g., $\hat{X}$) on the ellipse (dotted ellipse) has constant sum of distances from the speaker and microphone.

Figure 4 illustrates (a) the empty room impulse response and (b,c,d) the full responses with the pose kernels by varying the location and pose of an object. The left column shows the pose kernel $k_j(t)$ encoded to the physical space, while the right column shows the actual signal. Due to the fact that no bearing information is included from audio signal, each peak in the pose kernel $k_j(t)$ corresponds to a possible reflector location on the ellipse of which focal points coincide with the locations of the speaker and microphone.
with the spatial encoding of the pose kernel, we reformulate Equation (2):

\[
\{P_i(X)\}_{i=1}^N = g_\theta(\phi_v(X; I), \max_j \phi_a(K_j(X))),
\]

where \(\phi_v\) and \(\phi_a\) are the feature extractors for visual and audio signals, respectively.

Specifically, \(\phi_v\) is the visual features evaluated at the projected location of \(X\) onto the image \(I\), i.e.,

\[
\phi_v(X; I) = \{p_t(\Pi X)\}_{i=1}^N,
\]

where \(p_t \in [0, 1]^{W \times H}\) is the likelihood of the \(i^{th}\) landmark in the image \(I\). \(\Pi\) is the operation of 2D projection, i.e., \(p_t(\Pi X)\) is the likelihood of the \(i^{th}\) landmark at 2D projected location \(\Pi X\).

\(\phi_a(K_j(X))\) is the audio feature from the \(j^{th}\) pose kernel evaluated at \(X\). We use the max-pooling operation to fuse multiple received audio signals, which is agnostic to location and ordering of audio signals. This facilitates scene generalization where the learned audio features can be applied to a new scene with different audio configurations (e.g., the number of sources, locations, scene geometry).

We learn \(g_\theta\) and \(\phi_a\) by minimizing the following loss:

\[
\mathcal{L} = \sum_{I, K, P \in \mathcal{D}} \|g_\theta(\phi_v, \max_j \phi_a(K_j)) - \{\hat{P}_i\}_{i=1}^N\|_2^2, \tag{10}
\]

where \(\{\hat{P}_i\}_{i=1}^N\) is the ground truth 3D heatmaps, and \(\mathcal{D}\) is the training dataset. Note that this paper focuses on the feasibility of metric lifting by using audio signals where we use an off-the-shelf human pose estimator \(\{P_t\}_{i=1}^N [5]\).

### 3.3. Network Design and Implementation Details

We design a 3D convolution neural network (3D CNN) to encode 2D pose detection from an image (using OpenPose [5]) and four audio signals from microphones. Inspired by the design of the convolution pose machine [70], the network is composed of six stages that can increase the receptive field while avoiding the issue of the vanishing gradients. The 2D pose detection is represented by a set of heatmaps that are encoded in the 70 × 70 × 50 voxel grid via inverse projection, which forms 16 channel 3D heatmaps. For the pose kernel from each microphone, we spatially encode over a 70 × 70 × 50 voxel grid that are convolved with three 3D convolutional filters followed by max pooling across four audio channels. Each grid is 5 cm, resulting in 3.5 m × 3.5 m × 2.5 m space. These audio features are combined with the visual features to form the audio-visual features. These features are transformed by a set of 3D convolutions to predict the 3D heatmaps for each joint. The prediction, in turn, is combined with the audio-visual features to form the next stage prediction. The network architecture is shown in Figure 5.

We implemented the network with PyTorch, and trained it on a server using 4 Tesla v100 GPUs. SGD optimizer is used, and learning rate is 1. The model has been trained for 70 epochs (around 36 hours) until convergence.
Figure 5. We design a 3D convolutional neural network to encode pose kernels (audio) and 2D pose detection (image) to obtain the 3D metric reconstruction of a pose. We combine audio and visual features using a series of convolutions (audio features from multiple microphones are fused via max-pooling). The audio visual features are convolved with a series of 3 × 3 × 3 convolutional kernels to predict the set of 3D heatmaps for joints. We use multi-stage prediction, inspired by the convolutional pose machine architecture [70], which can effectively increase the receptive field while avoiding vanishing gradients.

Figure 6. We collect our PoseKernel dataset in different environments with at least six participants per location, totalling more than 6,000 poses.

4. PoseKernel Dataset

We collect a new dataset called PoseKernel dataset. It is composed of more than 10,000 frames of synchronized videos and audios from six locations including living room, office, conference room, laboratory, etc. For each location, more than six participants were asked to perform as shown in Figure 6.

The cameras, speakers, and microphones are spatially calibrated using off-the-shelf structure-from-motion software such as COLMAP [51] by scanning the environments with an additional camera and use the metric depth from the RGB-D cameras to estimate the true scale of the 3D reconstruction. We manually synchronize the videos and speakers by a distinctive audio signal, e.g., clapping, and the speakers and microphones are hardware synchronized by a field recorder (e.g., Zoom F8n Recorder) at a sample rate of 96 kHz.

For each scene, video data are captured by two RGB-D Azure Kinect cameras. These calibrated RGB-D cameras are used to estimate the ground truth 3D body pose using state-of-the-art pose estimation methods such as FrankMocap [49]. Multiple RGB-D videos are only used to generate the ground truth pose for training. In the testing phase, only a single RGB video is used.

Four speakers and four microphones are used to generate the audio signals. Each speaker generates a chirp audio signal sweeping frequencies between 19 kHz to 32 kHz. We use this frequency band because it is audible from consumer-grade microphones while inaudible for humans. Therefore, it does not interfere with human generated audio. In order to send multiple audio signals from four speakers, we use frequency-division multiplexing within the frequency band. Each chirp duration is 100 ms, resulting in 10 FPS reconstruction. At the beginning of every capture session, we capture the empty room impulse response for each microphone in the absence of humans.

We ask participants to perform a wide range of daily activities, e.g., sitting, standing, walking, and drinking, and range of motion in the environments. To evaluate the generalization on heights, in our test data, we include three minors (height between 140 cm and 150 cm) with consent from their guardians. All person identifiable information including faces are removed from the dataset.

5. Results

We evaluate our method on the PoseKernel dataset by comparing with state-of-the-art and baseline algorithms.

Evaluation Metric We use the mean per joint position error (MPJPE) and the percentage of correct keypoint (PCK) in 3D as the main evaluation metrics. For PCK, we report the PCKt where t is the error tolerance in cm.

Baseline Algorithms Three state-of-the-art baseline algorithms are used. (1) Lifting from the Deep, or Vis.:LfD [60] is a vision based algorithm that regresses the 3D pose from a single view image by learning 2D and
Table 1. We use MPJPE (the lower, the better) as the evaluation metric to compare our method with state-of-the-art vision based algorithms including LfD (Vis.:LfD [60]), FrankMocap (Vis.:Frank [49]), MeTro (Vis.:MeTro [50]) and our ablated algorithms including Audio×4 and Vision+Audio×2. We test on two sets: one with minor participants and one with adult participants (minor MPJPE/adult MPJPE). All numbers are reported in cm.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Head</th>
<th>Neck</th>
<th>Hand</th>
<th>Elbow</th>
<th>Shoulder</th>
<th>Hip</th>
<th>Knee</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vis.:LfD [60]</td>
<td>57.47</td>
<td>54.77</td>
<td>52.37</td>
<td>52.89</td>
<td>59.07</td>
<td>54.05</td>
<td>51.70</td>
<td>57.39</td>
</tr>
<tr>
<td>Vis.:Frank [49]</td>
<td>42.07</td>
<td>17.33</td>
<td>43.24</td>
<td>17.87</td>
<td>44.38</td>
<td>18.43</td>
<td>44.68</td>
<td>18.79</td>
</tr>
<tr>
<td>Vis.:MeTro [50]</td>
<td>85.37</td>
<td>66.24</td>
<td>89.93</td>
<td>65.91</td>
<td>97.80</td>
<td>71.17</td>
<td>93.18</td>
<td>69.80</td>
</tr>
<tr>
<td>Audio×4</td>
<td>313.4</td>
<td>290.0</td>
<td>321.2</td>
<td>259.1</td>
<td>323.6</td>
<td>265.0</td>
<td>143.3</td>
<td>143.8</td>
</tr>
</tbody>
</table>

Table 2. We use PCK@t (the higher, the better) as evaluation metric to compare our method with SOTA vision based algorithms including LfD (Vis.:LfD [60]), FrankMocap (Vis.:Frank [49]), MeTro (Vis.:MeTro [50]) and our ablated algorithms including Audio×4 and Vision+Audio×2. We test on two sets: one with minor participants and one with adult participants (minor PCK/adult PCK). All numbers are reported in cm.

<table>
<thead>
<tr>
<th>Methods</th>
<th>t</th>
<th>Head</th>
<th>Neck</th>
<th>Hand</th>
<th>Elbow</th>
<th>Shoulder</th>
<th>Hip</th>
<th>Knee</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vis.:LfD [60]</td>
<td>10 cm</td>
<td>0.001</td>
<td>0.068</td>
<td>0.008</td>
<td>0.152</td>
<td>0.000</td>
<td>0.051</td>
<td>0.002</td>
<td>0.055</td>
</tr>
<tr>
<td>Vis.:Frank [49]</td>
<td>10 cm</td>
<td>0.033</td>
<td>0.253</td>
<td>0.025</td>
<td>0.257</td>
<td>0.016</td>
<td>0.245</td>
<td>0.025</td>
<td>0.249</td>
</tr>
<tr>
<td>Vis.:MeTro [50]</td>
<td>10 cm</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Audio×4</td>
<td>10 cm</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Vis.+Audio×2</td>
<td>10 cm</td>
<td>0.590</td>
<td>0.409</td>
<td>0.541</td>
<td>0.515</td>
<td>0.410</td>
<td>0.224</td>
<td>0.459</td>
<td>0.270</td>
</tr>
<tr>
<td>Ours</td>
<td>10 cm</td>
<td>0.648</td>
<td>0.380</td>
<td>0.639</td>
<td>0.473</td>
<td>0.336</td>
<td>0.241</td>
<td>0.369</td>
<td>0.439</td>
</tr>
</tbody>
</table>

3D joint locations together. To resolve the depth ambiguity, a statistical model is learned to generate a plausible 3D reconstruction. This algorithm predicts 3D pose directly where we apply the Procrustes analysis to align with image projection. (2) FrankMocap (Vis.:Frank [49]) leverages the pseudo ground truth 3D poses on in-the-wild images that can be obtained by EFT [27]. Augmenting 3D supervision improves the performance of 3D pose reconstruction. This algorithm predicts the shape and pose using the SMPL parametric mesh model [35]. None of existing single view reconstruction approaches including these baseline methods produces metric scale reconstruction. Given their 3D reconstruction, we scale it to a metric scale by using the average human height in our dataset (1.7m). (3) MeTro (Vis.:MeTro [50]) offers metric scale reconstruction while the scale is data-driven instead of based on real signal level insights, thus heuristic. Note here both LfD and FrankMoCap did not include code for us to custom train it on our dataset. For the sake of fairness, we use pretrained model for all baselines. We believe the models are all trained on large enough datasets so that this comparison will not favor our own solution.

Our Ablated Algorithms In addition to the state-of-the-art vision based algorithms, we compare our method by ablating our sensing modalities. (1) Audio×4 uses four audio signals to reconstruct the 3D joint locations to study the impact of the 2D visual information. (2) Vis.+Audio×2 uses a single view image and two audio sources to predict the 3D joint location in the 3D voxel space. (3) Ours is equivalent to Vision+Audio×4. 5.1. PoseKernelLifter Evaluation Among the six environments in PoseKernel dataset, we use 4 environments for training and 2 environments for testing. The training data consists of diverse poses performed by six adult (whose heights range between 155 cm and 180 cm) and two minors (with heights 140 cm and 150 cm). The testing data includes two adult and one minor participants whose heights range between 140 cm and 180 cm. Comparison We measure the reconstruction accuracy using MPJPE metric summarized in Table 1. As expected, state-of-the-art vision based lifting approaches (Vis.:LfD and Vis.:Frank) that predict 3D human pose in a scale-free space are sensitive to the heights of the subjects, resulting in 18 ~ 40 cm mean error for adults and 40 ~ 60 cm for minors, respectively. Our results with vision and audio inputs show significant improvement in human pose reconstruction, with 4 cm mean error for both, adults and minors.
cm for minors, i.e., the error is larger for minor participants because their heights are very different from the average height 1.7 m. Vis.:Frank outperforms Vis.:LfD, because Vis.:Frank uses a larger training data, and thus estimates poses more accurately. Nonetheless, our pose kernel that is designed for metric scale reconstruction significantly outperforms these approaches. The performance is not dependent on the heights of participants. In fact, it produces around 20% smaller error for minor participants than the adult participants because of their smaller scale. Similar observations can be made in PCK summarized in Table 2.

Ablation Study We ablate the sensing components of our approach. As summarized in Tables 1 and 2, the 3D metric lifting leverage the strong cue from visual data. Without visual cue, i.e., Audio×4, the reconstruction is highly erroneous, while combining with audios as a complementary signal (Vis.+Audio×2 and Ours) significantly improve the accuracy. While providing metric information, reconstructing the 3D human pose from audio signals alone (Audio×4) is very challenging because the signals are (1) non-directional: a received signal is integration of audio signals over all angles around the microphone which does not provide bearing angle unlike visual data; (2) non-identifiable: the reflected audio signals are not associated with any semantic information, so it is difficult to tell where a specific reflection is coming from; and (3) slow: due to the requirement of linear frequency sweeping (10 Hz), the received signals are blurred in the presence of body motion, which is equivalent to an extremely blurry image created by 100 ms exposure with a rolling shutter. Nonetheless, augmenting audio signals improve 3D metric reconstruction regardless the heights of the participants.

Generalization We report the results in completely different testing environments, which show the strong generalization ability of our method. For each environment, the spatial arrangement of the camera and audio/speakers is different, depending on the space configuration. Figure 7 visualizes the qualitative results of our 3D pose lifting method where we successfully recover the metric scale 3D pose in different environments. We also include a failure cases in the presence of severe occlusion as shown in Figure 7(d).
References


[34] Adrian Llopart. Liftformer: 3d human pose estimation using attention models. arXiv, 2020. 1, 2


[54] Sheng Shen, He Wang, and Romit Roy Choudhury. I am a smartwatch and i can track my user’s arm. In Annual International Conference on Mobile Systems, Applications, and Services, 2016. 3


[56] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In ECCV, 2018. 2

[57] Camillo J Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. Computer Vision and Image Understanding, 2000. 2


[60] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In CVPR, 2017. 1, 2, 6, 7


