# Human Pose Estimation from an Extremely Low-Resolution Image Sequence by Pose Transition Embedding Network

Yasutomo Kawanishi[1] [a], Hitoshi Nishimura[2] [b] and Hiroshi Murase[3] [c]

[1]*Guardian Robot Project, RIKEN, Kyoto, Japan*
[2]*KDDI Research, Saitama, Japan*
[3]*Graduate School of Informatics, Nagoya University, Aichi, Japan*

Keywords: Low Resolution, Human Pose Estimation, Temporal Information.

Abstract: This paper addresses the problem of human pose estimation from an extremely low-resolution (ex-low) image sequence. In an ex-low image (e.g., $16 \times 16$ pixels), it is challenging, even for human beings, to estimate the human pose smoothly and accurately only from a frame because of resolution and noise. This paper proposes a human pose estimation method, named Pose Transition Embedding Network, that considers the temporal continuity of human pose transition by using a pose-embedded manifold. This method first builds a pose transition manifold from the ground truth of human pose sequences to learn feasible pose transitions using an encoder-decoder model named Pose Transition Encoder-Decoder. Then, an image encoder, named Ex-Low Image Encoder Transformer, encodes an ex-low image sequence into an embedded vector using a transformer-based network. Finally, the estimated human pose is reconstructed using a pose decoder named Pose Transition Decoder. The performance of the method is confirmed by evaluating an ex-low human pose dataset generated from a publicly available action recognition dataset.

## 1 INTRODUCTION

Human pose estimation is an essential task in various computer vision applications such as action recognition (Song et al., 2021), motion prediction (Fujita and Kawanishi, 2024), anomaly detection (Temuroglu et al., 2020), and internal state recognition (Mizuno et al., 2023). It has been applied to various kinds of video, such as in-vehicle cameras, drone cameras, smartphone cameras, and surveillance cameras. Because of their importance, this topic has been actively developed, and state-of-the-art methods have achieved very high accuracy, even in complicated scenes. One of the most important applications of human pose estimation is skeleton-based human action recognition and prediction. The estimated results should be accurate enough when using the human pose estimation results in pose-based action recognition tasks; the estimated results should be accurate enough. In addition, they should be temporally as smooth as the actual human poses to express their hu-

man motion.

Most existing human pose estimation methods require the persons in an image to be somewhat large, for example, more than 100 pixels in height. However, in videos for practical applications, such as in-vehicle cameras or surveillance cameras, the size of the person is often small in the frame of the videos, that is, person images are often of low resolution. Even with recent advances in camera sensors, the size of persons captured by cameras from afar remains small. If given an extremely low-resolution cropped image (ex-low; e.g., a person is $16 \times 16$ pixels), is it possible to estimate the human pose from the ex-low input? In this study, we focus on a situation in which the size of a person in a cropped image is very small.

If a person image is extremely low-resolution (*ex-low*), human pose estimation becomes difficult for the following reasons. First, an ex-low image contains little information for estimating human pose. It is also difficult to distinguish the body region of the target person in an image from the background of the image because of the poor features in the ex-low image and blurry boundaries. Because the number of pixels for a person in an ex-low image is small, the signal-to-noise ratio is low and the effect of salt-and-

[a] https://orcid.org/0000-0002-3799-4550
[b] https://orcid.org/0000-0002-9552-3837
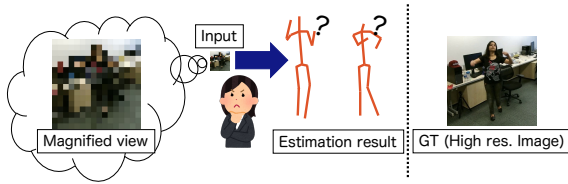[c] https://orcid.org/0000-0002-8103-9294

Figure 1: Human pose estimation from an ex-low image is difficult, even for humans, because the input contains less information. The small image in the middle represents the input. The image on the left shows a magnified view of the input. A high-resolution version of the input is shown on the right side (for reference). (These photos are originally from the NTU RGB+D dataset (Shahroudy et al., 2016).)
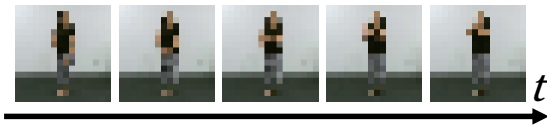


Figure 2: Temporal information helps us estimate human pose. We think you can guess how the person is moving.

pepper noise is relatively more significant. These issues make pose estimation from an ex-low image difficult. As shown in Fig. 1, it is extremely difficult even for humans to use only a single ex-low image.

Meanwhile, once we see a video (i.e., an ex-low image sequence), we can guess how the human pose changes (Fig. 2). This implies that temporal information is powerful for pose estimation. This is because human pose transitions have temporal continuity. In this study, we focus on the temporal continuity of human pose transition and propose a pose estimation method named Pose Transition Embedding Network from an extremely low-resolution (ex-low) image sequence. Human detection and tracking should be applied beforehand to handle the image sequence of the target person. Thus, a top-down human pose estimation approach that estimates the human pose after human detection is suitable for this scenario. This study assumes that each human region is detected and tracked during preprocessing.

To address this pose estimation problem, we propose Pose Transition Embedding Network to estimate an accurate and smooth human pose sequence from an ex-low image sequence. This method consists of two parts: the Pose Transition Encoder-Decoder, which captures how the human pose changes, and Ex-Low Image Encoder Transformer, which extracts a feature from an ex-low image sequence.

First, the Pose Transition Encoder-Decoder model is trained to capture the continuity of the human pose transition. Because there is a strong correlation between human poses in adjacent frames, pose transition can be described as a feature vector in a low-dimensional embedding space. The encoder-decoder

model is trained using ground-truth pose annotation sequences to encode human pose transitions into vectors in a low-dimensional embedding space. After training, each vector in the space is associated with a feasible human-pose transition.

Then, the Ex-Low Image Encoder Transformer, followed by the Pose Transition Decoder, learns the mapping from an input image sequence to the human pose sequence. The Ex-Low Image Encoder Transformer captures the spatial and temporal variations of the input ex-low image sequence using CNN and Transformer structures. The reconstructed human-pose sequence is expected to be smooth and feasible.

The contributions of this paper are summarized as follows;

- We addressed a new computer vision problem of human pose estimation from an extremely low-resolution (ex-low; $16 \times 16$ pixels) image sequence.

- We propose the *Pose Transition Embedding Network*, which consists of the *Pose Transition Encoder-Decoder model* and the *Ex-Low Image Encoding Transformer*. This method can handle the feasible temporal transitions of human poses in an embedded space.

- We also propose a pseudo dataset generation method based on the existing datasets.

The remainder of this paper is organized as follows. In Section 2, recent studies on human pose estimation are summarized. In Section 3, the details of the proposed Pose Transition Embedding Network are described. In Section 4, experimental results are presented. Finally, we conclude the paper in Section 5.

## 2 RELATED WORK

In this section, we first summarize single-frame human pose estimation methods that have been actively developed in computer vision. Then, because we focus on low-resolution images, we summarize the human pose estimation methods for low-resolution images.

### 2.1 Human Pose Estimation

Human pose estimation has been widely developed and applied to various applications. Generally, human pose estimation methods can be divided into two categories: top-down (e.g., (Xiao et al., 2018)) and bottom-up approaches (e.g., (Kreiss et al., 2019)). The top-down approach first detects human bounding boxes, and then estimates the human pose for each

bounding box. Generally, the detected human bounding boxes are resized to fit the input for the pose estimator. This approach is robust to the size of the target person, because the input to the pose estimator is resized. DeepPose (Toshev and Szegedy, 2014) is the most earliest deep learning-based human pose estimation. This method directly estimates the joint locations of the human body using a regression model.

Recently, heatmap-based approaches have been mainly used for human pose estimation. This approach first estimates the heat maps of the body joints, and then selects the actual locations from the heat maps. PoseNet (Papandreou et al., 2017a) estimates the heatmaps of the body joints and their offset maps. While state-of-the-art methods have become complicated, a Simple Baseline (Xiao et al., 2018) achieves good performance even with a very simple network. This method is widely used as a baseline for top-down human pose estimation.

These methods assume that only one person is present in an input image. To handle temporal continuity, the target person must be tracked. In this case, a sequence of bounding boxes of the target person is provided as an input to the pose estimation process. Therefore, this top-down approach makes it easier to extend the multi-frame human pose estimation.

Meanwhile, recent studies focus on bottom-up human pose estimation. This approach first estimates the heatmap of each body joint and then finds an optimal combination of their locations. Several well-known methods are available, including OpenPose (Cao et al., 2021), and PifPaf (Kreiss et al., 2019). This approach is weak for low-resolution human pose estimation because each body joint of the people small in the image becomes too small.

## 2.2 Multi-Frame Human Pose Estimation

Several studies have been proposed to handle temporal information for human pose estimation. A straightforward approach is to use convolutional LSTMs, such as LSTM Pose Machine (Luo et al., 2018), UniPose (Artacho and Savakis, 2020), Motion Adaptive Pose Esitmation (Fan et al., 2021), and FAMI Pose (Liu et al., 2022). (Liu et al., 2022). Another approach involves estimating the motion of a target human body as an intermediate representation (Liu et al., 2021). This approach includes Flowing ConvNets (Pfister et al., 2015) and Thin-Slicing Network (Song et al., 2017). These methods estimate motion flows to incorporate information from adjacent frames for pose estimation. DCPose (Liu et al., 2021) estimates motion offsets using Pose Residual

Fusion. The above approach is also adopted as a heatmap-based approach.

## 2.3 Estimation from Low-Resolution Images

The accuracy of human pose estimation using a heatmap-based approach is limited by the resolution of the output heatmap (the same as that of the input image). To tackle the difficulty of the heatmap-based approach, several researchers have proposed the offset-map-based human pose estimation methods (Papandreou et al., 2017b; Zhang et al., 2019). These methods output an offset map for each body joint. Each pixel value in the offset map indicates the offset of the target keypoint from the pixel. They also output a binary heatmap for each body joint and calculate each keypoint location by averaging the offset map values within the selected pixels in the corresponding binary heatmap.

These methods increase the pose-estimation accuracy even if the input image is low resolution. Wang et al. (Wang et al., 2022) extended this method by replacing a binary heatmap with a Gaussian distribution. The method was evaluated using a low-resolution version of the MSCOCO (Lin et al., 2014) dataset with a resolution of $128 \times 96$ pixels.

Srivastav et al. (Srivastav et al., 2019) have proposed a human pose estimation from a low-resolution depth image. They use low-resolution images for privacy protection since their target situation is medical surgery where there are several medical doctors and a patient. They used a low-resolution version of the MVOR dataset (Srivastav et al., 2018). In the paper, the resolution of the input images is $64 \times 48$ pixels. The method learns super-resolution and bottom-up pose estimation simultaneously.

Xu et al. (Xu et al., 2020) have proposed the RSC-Net, which can estimate 3D human pose and shape from a low-resolution image. In the paper, the resolution of the input image is $32 \times 32$ pixels. They parametrize the 3D model of a person by using SMPL model (Loper et al., 2015). The model is trained using multi-scale images, not only low-resolution images.

Iwata et al. (Iwata et al., 2021) introduced LFIR2Pose to estimate the human pose from $16 \times 16$ Far-infrared (LFIR) image sequence, which makes it easier to distinguish the traget person from the background. By assuming that only one person is in an LFIR image, they estimate the human pose based on the top-down approach. The model is a 3D Convolutional Neural Network followed by a regression network. This method is very simple, but effectively uses temporal information for human pose estimation
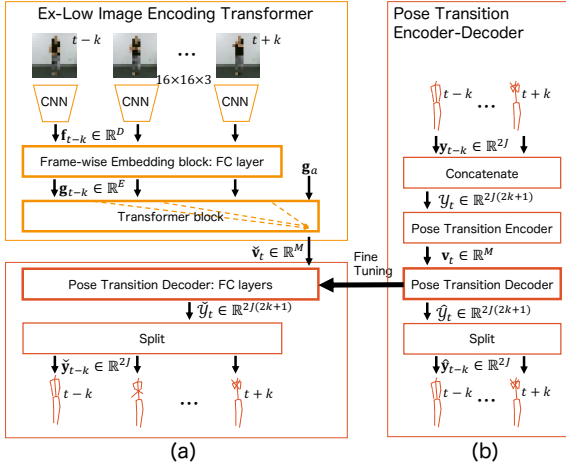
Figure 3: (a) The proposed Temporal Embedding Network consists of the Ex-Low Image Encoding Transformer followed by the Pose Transition Decoder. The texts beside the arrows in the figure indicate the dimension of the data. (b) Pose Transition Encoder-Decoder model. The texts beside the arrows in the figure denote the dimension of the data.

from a low-resolution image sequence. In LFIR images, the human body and background can be distinguished easily because the temperature of the human body is relatively higher than that of the background in a room. This is an advantage against using RGB images; however, the method would not work outside, especially under the sun.

# 3 POSE TRANSITION EMBEDDING NETWORK

## 3.1 Overview

Human pose estimation from an ex-low image sequence is difficult because of the lack of information in the image, the ambiguity of the border between a target person, cluttered background, and the effect of noise. The proposed method overcomes these difficulties by focusing on the temporal continuity and feasible transitions of human poses. The input of the method is an ex-low image sequence $I_t = (I_{t-k}, \ldots, I_t, \ldots, I_{t+k})$, that is, $2k + 1$ frames around frame $t$, and the output is the human pose $\mathbf{y}_t$ in the middle image $I_t$. This section describes the proposed method for extremely low-resolution (ex-low) human pose estimation, named Temporal Embedding Network, which consists of the Pose Transition Encoder-Decoder and the Ex-Low Image Encoding Transformer.

## 3.2 Ex-Low Image Encoding Transformer

This model captures the temporal information of the human pose sequence from an ex-low image sequence using a convolutional neural network and transformer. The CNN captures spatial information from each frame, and then the transformer captures temporal information. Finally, the Ex-Low Image Encoding Transformer outputs an embedded vector $\check{\mathbf{v}}_t \in \mathbb{R}^M$. The architecture of the model is visualized in Fig. 3 (a).

First, each color image $I_t \in \mathbb{R}^{16 \times 16 \times 3}$ in the input sequence $I_t = (I_{t-k}, \ldots, I_t, \ldots, I_{t+k})$ is fed into a CNN model specialized for an ex-low image, and each feature vector $\mathbf{f}_t \in \mathbb{R}^D$ corresponds to the input image $I_t$ is obtained. Then, each of the features is embedded into $E$ dimensional space using a Multi-Layer Perceptron (MLP) layer, and each feature vector $\mathbf{g}_t \in \mathbb{R}^E$ is obtained. This sequence of features together with an feature aggregation token $\mathbf{g}_a$ is fed into the transformer block, which consists of multiple transformer layers. Each transformer layer consists of a multi-head attention layer and an MLP layer. In this paper, the number of layers was determined empirically and set to four. Among the outputs of the transformer layers, the aggregated output $\check{\mathbf{v}}_t \in \mathbb{R}^M$ is selected as the final output of this module. This procedure is denoted as

$$\check{\mathbf{v}}_t = f_i(I_t). \tag{1}$$

The embedded vector $\check{\mathbf{v}}_t$ is fed to the Pose Transition Decoder $f_d$ explained in Section 3.3, and a pose sequence $\check{\mathcal{Y}}_t = (\check{\mathbf{y}}_{t-k}, \ldots, \check{\mathbf{y}}_t, \ldots, \check{\mathbf{y}}_{t+k})$ is obtained as

$$\check{\mathcal{Y}}_t = f_d(\check{\mathbf{v}}_t) = f_d(f_i(I_t)). \tag{2}$$

We empirically define the function $f_i$ by four $3 \times 3$ CNN layers whose channels are 16, 32, 64, and 128, followed by a $1 \times 1$ convolution to obtain channel 64.

## 3.3 Pose Transition Encoder-Decoder

This model aims to capture temporally smooth and feasible human pose transitions based on an encoder-decoder architecture. The decoder part is used for pose estimation by combining it with the Ex-Low Image Encoding Transformer.

Here, we assume that human pose transition can be described in a low-dimensional space. The encoder encodes a human pose sequence into a low-dimensional vector, and then the decoder decodes the feasible human pose sequence. The architecture of the model is visualized in Fig. 3 (b).

Human pose is described as a set of body joint locations. Each body joint location in a 2D coordinate system is described as a two-dimensional vector.

Therefore, the human pose at frame $t$ can be described as a $2J$ dimensional vector $\mathbf{y}_t \in \mathbb{R}^{2J}$, where $J$ denotes the number of body joints. Thus, a pose transition consisting of $2k+1$ frames around time $t$ can be described as a concatenated vector $\mathcal{Y}_t \in \mathbb{R}^{2J(2k+1)}$.

Here, we assume that the pose transition is restricted such that they form a low-dimensional manifold in the $M$-dimensional pose-transition space. We named this low-dimensional manifold *Pose Transition Manifold*. The proposed model estimates an embedded vector $\mathbf{v}_t \in \mathbb{R}^M$ from input $\mathcal{Y}_t$ using an AutoEncoder whose intermediate dimension is $M$. The encoder and decoder of the AutoEncoder are denoted as $f_e$ and $f_d$, respectively.

$$\mathbf{v}_t = f_e(\mathcal{Y}_t), \qquad (3)$$
$$\widehat{\mathcal{Y}}_t = f_d(\mathbf{v}_t). \qquad (4)$$

The details of $f_e(\cdot)$ and $f_d(\cdot)$ are visualized in Fig. 3. Both of them consist of two fully connected layers. This encoder-decoder model is trained to reduce the reconstruction loss, $L_r$, which is defined as the sum of the Euclidean distances of all body joints. Its equation is as follows,

$$L_r = \sum_{(\mathbf{y}_i, \widehat{\mathbf{y}}_i, \mathbf{m}_i) \in (\mathcal{Y}_t, \widehat{\mathcal{Y}}_t, \mathcal{M}_t)} d(\mathbf{y}_i, \widehat{\mathbf{y}}_i, \mathbf{m}_i), \qquad (5)$$

$$d(\mathbf{a}, \mathbf{b}, \mathbf{m}) = \sum_{j=1}^{J} m_j \sqrt{(a_{2j-1} - b_{2j-1})^2 + (a_{2j} - b_{2j})^2}, \qquad (6)$$

where $\mathbf{m}_i \in \{0,1\}^J$ is a mask indicating visible body joints in frame $i$, and $\mathcal{M}_t$ is a sequence of $\mathbf{m}_i$ around a frame at $t$. Note that $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$ are considered as sequences of $\mathbf{y}_i$ and $\widehat{\mathbf{y}}_i$ in equation (5) around a frame at $t$, respectively.

We empirically set $M = 40$, and the functions $f_e$ and $f_d$ using two fully-connected layers with a middle layer dimension of 85.

## 3.4 Training the Whole Model

The proposed Pose Transition Embedding Network is trained in an End-to-End manner with a pre-trained Pose Transition Encoder-Decoder.

First, the Pose Transition Encoder-Decoder is trained using the ground truth pose sequence to reconstruct the input themselves using equation (5). Then, the decoder part of the model is extracted and connected to the Ex-Low Image Encoding Transformer to build the Temporal Embedding Network as shown in equation (2). Finally, the model is trained in a supervised learning manner, using input image sequences and corresponding ground truth pose sequences by
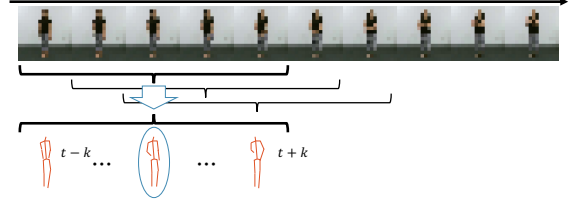


Figure 4: Temporal sliding window approach with stride one frame for long sequence. The estimated result of the center frame in a sliding window is selected for the result of the corresponding frame.

minimizing the pose estimation loss defined as,

$$L_e = \sum_{(\mathbf{y}_i, \check{\mathbf{y}}_i, \mathbf{m}_i) \in (\mathcal{Y}_t, \check{\mathcal{Y}}_t, \mathcal{M}_t)} d(\mathbf{y}_i, \check{\mathbf{y}}_i, \mathbf{m}_i). \qquad (7)$$

Note that the Pose Transition Decoder is fine-tuned in the training.

## 3.5 Pose Estimation for a Long Sequence

We use a temporal sliding window approach to estimate human poses in a long sequence. The proposed model accept $2k+1$ frames of ex-low images $I_t$ around time $t$, and outputs $2k+1$ poses $(\check{\mathbf{y}}_{t-k}, \ldots, \check{\mathbf{y}}_t, \ldots, \check{\mathbf{y}}_{t+k})$. For the final estimation result, we select the center of the output $\check{\mathbf{y}}_t$ as shown in Fig. 4. We apply this sliding window with a stride of one frame.

# 4 EVALUATION

## 4.1 Dataset

To evaluate ex-low human pose estimation, we require a dataset consisting of low-resolution human images. As the proposed method utilizes temporal information, the input should be a sequence of ex-low images. In addition, the dataset should contain diverse poses. Therefore, we generated a dataset from a large-scale video action recognition dataset.

In this evaluation, we selected the NTU RGB+D (Shahroudy et al., 2016) dataset as the source dataset. This dataset consists of videos captured by Kinect v2 sensors. It contains 60 human action classes acted on by 40 participants. The resolution of the images is $1,920 \times 1,080$ pixels. This dataset also has 2D/3D skeletons data provided by the Kinect sensors. Since the 2D/3D skeleton data is not very accurate, we applied YOLOv8-pose (Jocher et al., 2023) to estimate 2D human poses for each high-resolution frame. Because the pose estimation
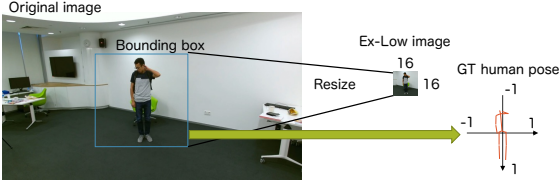
Figure 5: Generation of the dataset. The human pose and bounding box were estimated using YOLOv8-pose. Each cropped image is resized to $16 \times 16$ pixels. Each human pose is normalized to the $[-1, 1]$ range in the cropped image.

on high-resolution images by YOLOv8-pose is quite accurate, we use them as the ground truth human poses. Within the pose estimation results, we only use reliable samples based on the confidence scores of the pose estimator. Note that the "ground truth" in this case is, in fact, a silver standard.

We cropped a human body from each image with a square bounding box, and they were resized to $16 \times 16$ pixels ex-low images. Each ground-truth human pose was normalized to the $[-1, 1]$ range within the corresponding bounding box. In this coordinate, the height of each bounding box was 2, thus the value of 1 in the corrdinate can be considered to be approximately 0.8 m. The procedure is illustrated in Fig. 5. For the dataset, we selected $2k + 1$ consecutive frames with one stride, where one human pose can be estimated from an image. In this experiment, we used $k = 2$; thus the length of the sequence was five frames.

We divided the 40 subjects into five groups for five-fold cross-validation. In each split, we used 29 subjects for training, 3 subjects for validation, and 8 subjects for testing.

## 4.2 Evaluation Metrics

We evaluated the results from two viewpoints: the accuracy of the estimation and the smoothness of the estimation as the ground truth. For accuracy, we used the average of the Euclidean distance between the corresponding body joints. We named this metric the Independent Frame Error (*IFE*) in this paper, defined as

$$IFE(\mathcal{A}, \mathcal{B}, \mathcal{M}) = \frac{1}{2k+1} \sum_{(\mathbf{a}_i, \mathbf{b}_i, \mathbf{m}_i) \in (\mathcal{A}, \mathcal{B}, \mathcal{M})} d_m(\mathbf{a}_i, \mathbf{b}_i, \mathbf{m}_i), \tag{8}$$

$$d_m(\mathbf{a}, \mathbf{b}, \mathbf{m}) = \frac{1}{\sum_{j=1}^{J} m_j} d(\mathbf{a}, \mathbf{b}, \mathbf{m}), \tag{9}$$

where $\mathcal{A}$, $\mathcal{B}$, $\mathcal{M}$ are estimated, ground-truth, and mask sequences, respectively.

On the other hand, we used the absolute difference of inter-frame differences of the corresponding body joints between the estimated and ground-truth sequences as a smoothness metric. This is based on
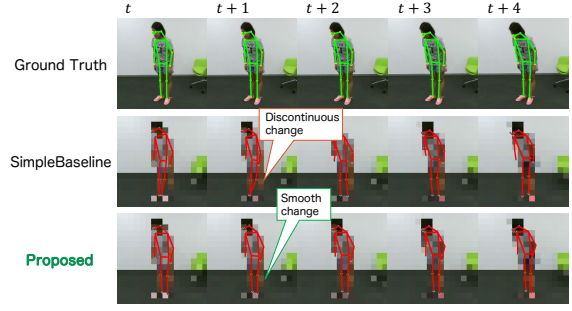


Figure 6: Example of pose estimation results. We can see both methods quite well estimate the human poses; but the results of SimpleBaseline is unstable around the ground-truth body-joint locations, which makes them non-smooth.

the idea that the estimated data should be as smooth as the ground truth. We named this metric the Frame Difference Absolute Error (*FDAE*). First, this metric calculates the inter-frame difference in the Euclidean distance between adjacent frames and then calculates their absolute difference from that of the ground truth for each frame. The FDAE is calculated as follows,

$$FDAE(\mathcal{A}, \mathcal{B}, \mathcal{M}) = \frac{1}{2k} \sum_{i=1}^{2k} d_e(\mathcal{A}, \mathcal{B}, \mathcal{M}, i), \tag{10}$$

$$d_e(\mathcal{A}, \mathcal{B}, \mathcal{M}, i) = |d_f(\mathcal{A}, \mathcal{M}, i) - d_f(\mathcal{B}, \mathcal{M}, i)|, \tag{11}$$

$$d_f(\mathcal{A}, \mathcal{M}, i) = d_m(\mathbf{a}_i, \mathbf{a}_{i+1}, \mathbf{m}_i^{i+1}), \tag{12}$$

$$\mathbf{m}_i^{i+1} = \mathbf{m}_i \odot \mathbf{m}_{i+1}, \tag{13}$$

where $\odot$ is the Hadamard product of vectors. Equation (12) calculates the inter-frame difference between $i$-th and $i+1$-th frames using Equation (6) with a mask, which is an intersection of the masks of the frames.

Additionally, We also use mean Average Precision (mAP) of each joint based on the Object Keypoint Similarity (OKS) defined in COCO Keypoint Detection Task (Lin et al., 2015).

## 4.3 Comparison with Existing Method and Ablation Study

We compared the proposed method with top-down and bottom-up methods. For the bottom-up methods, we just applied well-known pre-trained methods to the ex-low images (OpenPifPaf (Kreiss et al., 2019)). Also, for the top-down method, we applied Pose ResNet, as known as SimpleBaseline (Xiao et al., 2018) trained on our dataset. The method is a simple, but known to be able to provide a strong baseline. Since this method estimate human pose from an image one by one, it does not use temporal information.

Table 1: Pose estimation results. We compared the proposed method with the existing bottom-up and top-down methods. Most bottom-up methods cannot detect any pose from a low-resolution image. An ablation study was also conducted.

| Method | | IFE $\downarrow$ | FDAE $\downarrow$ | mAP (%) $\uparrow$ |
|---|---|---|---|---|
| Existing | Bottom-up method (OpenPifPaf (Kreiss et al., 2019)) | inappricable | | 0 |
| | Top-down method (SimpleBaseline (Xiao et al., 2018)) | 0.1076 | 0.0879 | 78.142 |
| Proposed | *IndependentCNN* (= w/o temporal information) | 0.0944 | 0.0573 | 82.802 |
| | *ChannelCombinedCNN* (= w/o Transformer) | 0.0935 | 0.0494 | 83.158 |
| | *CNNTransformer* (= w/o Pre-training) | 0.0789 | 0.0466 | 88.279 |
| | **Full model** | **0.0789** | **0.0466** | **88.314** |

As described in Section 3, the proposed method consists of a CNN, Transformer, and AutoEncoder. As an ablation study, we also compared with the three methods ablated from the proposed method; no temporal information (*Independent CNN*), temporal information (*ChannelCombinedCNN*), and temporal information with the transformer module (*CNNTransformer*. This can be considered as the proposed method without pretraining of the Pose Transition Decoder.), while the proposed method utilize pretrained autoencoder for decoding the pose sequence.

## 4.4 Experimental Results

The results are summarized in Table 1. Because it is very difficult to detect small body parts from ex-low images, most bottom-up methods cannot detect any poses, while the top-down methods can somehow detect poses. It is because top-down methods assume that there is a person in an image. By comparing the top-down method, which is a heatmap-based method, with IndependentCNN, we can see the heatmap-based is not suitable for this ex-low task. By comparing ChannelCombinedCNN and CNNTransformer, we can see spatio-temporal attention in the transformer network contribute to smooth and accurate estimation. The methods that use temporal information achieved lower score in the FDAE. In addition, from the table, we can see that the transformer can help improve accuracy. In this evaluation, the full model slightly outperformed the CNNTransformer. It is because CNNTransformer can also capture temporal transition quite well.

## 5 CONCLUSION

This study addressed the problem of human pose estimation from an extremely low-resolution (ex-low) image sequence. From an application perspective, the estimated human pose must be accurate and temporally smooth. This paper proposes a human pose estimation method, named the Pose Transition Embedding Network, that considers the temporal continuity of human pose transition by using a pose-embedded manifold. The Ex-Low Image Encoding Transformer captures spatial and temporal information and embeds them into a feature vector. The Pose Transition Decoder then reconstructs the feasible human pose from the feature vector. The evaluation results demostrated that the proposed method can estimate accurate and smooth poses.

Analyzing and optimizing the architecture of the network will be the subject of future work. In addition, the method was evaluated only on cropped images from the NTU RGB+D dataset. The dataset contains several scenes and multiple people; however, it would be better to evaluate various datasets. In addition, the top-down approach requires human body detection before pose estimation. In future work, we will develop a method for ex-low human body detection.

## ACKNOWLEDGEMENTS

## REFERENCES

Artacho, B. and Savakis, A. (2020). UniPose: Unified human pose estimation in single images and videos. In *Proc. 2020 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 7033–7042.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 43(01):172–186.

Fan, Z., Liu, J., and Wang, Y. (2021). Motion adaptive pose estimation from compressed videos. pages 11699–11708.

Fujita, T. and Kawanishi, Y. (2024). Recurrent graph convolutional network for sequential pose prediction from 3D human skeleton sequence. In *Proc. 27th International Conf. on Pattern Recognit.*, pages 342–358.

Iwata, S., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H., and Aizawa, T. (2021). LFIR2Pose: Pose estimation from an extremely low-resolution fir image sequence. In *Proc. 25th International Conf. on Pattern Recognit.*, pages 2597–2603.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics YOLO. (accessed on January 26, 2025).

Kreiss, S., Bertoni, L., and Alahi, A. (2019). PifPaf: Composite fields for human pose estimation. In *Proc. 2019 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 11969–11978.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common objects in context. arXiv:1405.0312.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV2014*, pages 740–755.

Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S., Yang, B., and Wang, X. (2021). Deep dual consecutive network for human pose estimation. In *Proc. 2021 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 525–534.

Liu, Z., Feng, R., Chen, H., Wu, S., Gao, Y., Gao, Y., and Wang, X. (2022). Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proc. 2022 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 11006–11016.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16.

Luo, Y., Ren, J., Wang, Z., Sun, W., Pan, J., Liu, J., Pang, J., and Lin, L. (2018). LSTM pose machines. In *Proc. 2018 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 5207–5215.

Mizuno, M., Kawanishi, Y., Fujita, T., Deguchi, D., and Murase, H. (2023). Subjective baggage-weight estimation from gait: Can you estimate how heavy the person feels? In *Proc. 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 567–574.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., and Murphy, K. (2017a). Towards accurate multi-person pose estimation in the wild. In *Proc. 2017 IEEE Conf. on Comput. Vision and Pattern Recognit.*, pages 3711–3719.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., and Murphy, K. (2017b). Towards

accurate multi-person pose estimation in the wild. In *Proc. 2017 IEEE Conf. on Comput. Vision and Pattern Recognit.*, pages 3711–3719.

Pfister, T., Charles, J., and Zisserman, A. (2015). Flowing ConvNets for human pose estimation in videos. In *Proc. 15th International Conf. on Comput. Vision*, pages 1913–1921.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proc. 2016 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 1010–1019.

Song, J., Wang, L., Van Gool, L., and Hilliges, O. (2017). Thin-Slicing Network: A deep structured model for pose estimation in videos. In *Proc. 2017 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 5563–5572. IEEE.

Song, L., Yu, G., Yuan, J., and Liu, Z. (2021). Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055.

Srivastav, V., Gangi, A., and Padoy, N. (2019). Human pose estimation on privacy-preserving low-resolution depth images. In *Proc. 22nd Medical Image Computing and Computer Assisted Intervention*, pages 583–591.

Srivastav, V., Issenhuth, T., Kadkhodamohammadi, A., de Mathelin, M., Gangi, A., and Padoy, N. (2018). MVOR: A multi-view rgb-d operating room dataset for 2D and 3D human pose estimation. In *Proc. 2018 MICCAI Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*.

Temuroglu, O., Kawanishi, Y., Deguchi, D., Hirayama, T., Ide, I., Murase, H., Iwasaki, M., and Tsukada, A. (2020). Occlusion-aware skeleton trajectory representation for abnormal behavior detection. In *Proc. 26th International Workshop on Frontiers of Computer Vision*, volume 1212, pages 108–121, Singapore. Springer Singapore.

Toshev, A. and Szegedy, C. (2014). DeepPose: Human pose estimation via deep neural networks. In *Proc. 2014 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit.*, pages 1653–1660.

Wang, C., Zhang, F., Zhu, X., and Ge, S. S. (2022). Low-resolution human pose estimation. *Pattern Recognition*, 126:108579.

Xiao, B., Wu, H., and Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In *Computer Vision – ECCV2018*, volume 11210, pages 472–487.

Xu, X., Chen, H., Moreno-Noguer, F., Jeni, L. A., and De la Torre, F. (2020). 3D human shape and pose from a single low-resolution image with self-supervised learning. In *Computer Vision – ECCV2020*, volume 12354, pages 284–300.

Zhang, R., Zhu, Z., Li, P., Wu, R., Guo, C., Huang, G., and Xia, H. (2019). Exploiting offset-guided network for pose estimation and tracking. In *Proc. 2019 IEEE/CVF Conf. on Comput. Vision and Pattern Recognit. Workshops*, pages 1–9.