RF-based End-to-End Multi-Person Pose Estimation using Visual Clue

Seunghyun Kim, Seunghwan Shin, Sangwon Lee, Yusung Kim*

Sungkyunkwan Univiersity
kim95175@skku.edu, @skku.edu, kl0081kl@g.skku.edu, yskim525@skku.edu

Abstract

Compared to the traditional camera-based methods, radio frequency (RF) based pose estimation has great potential to be used when the field of vision is obstructed. We present an RF-based Pose Estimation framework with Transformer (RPET), as the first RF-based method operating in a fully end-to-end fashion with an easy-to-install portable radar. Furthermore, RPET does not require complicated pre-processing works, and hand-crafted post-processing modules such as RoI cropping, NMS, and keypoint grouping. We also introduce a novel idea called Visual Clue (VC), which mimics an image feature map represented in a camera-based method. The extensive experimental results demonstrate that VC can be extracted from RF signals, which improves representation learning for predicting multi-person pose estimation. We also show that RPET can generalize learning, as shown in situations where obstacles, that are unexposed during training, were added later or when tested in different locations.

Introduction

The study for multi-person pose estimation (MPPE) is one of the main research areas in the field of computer vision. It has achieved great performance improvement (Wang et al. 2021; Tian, Chen, and Shen 2019) along with the development of deep learning models such as CNN (He et al. 2016; Simonyan and Zisserman 2015) and Transformer (Vaswani et al. 2017). However, camera-based methods have limitations when the field of vision is physically blocked or is impaired due to dark lighting. Recently, RF-based methods have been studied to predict human poses using reflected RF signals (Zhao et al. 2018a; Wang et al. 2019; Zheng et al. 2022). RF signals are not affected by brightness, and can pass through obstacles such as walls. These characteristics work as advantages in military, security, disaster relief and other situations where visibility is not guaranteed. In addition, RF-based methods (Yue et al. 2020) may be preferred over camera-based methods because of privacy concerns, such as when continuous behavioral anomaly detection is needed (e.g. elderly care).

RF-based methods also pose several challenges. The first challenge is complexity. Assuming 2D pose estimations, RF-based methods (Adib et al. 2014; Zhao et al. 2018a; Zheng et al. 2022) usually require special radar hardware and complex pre-processing works to be performed on the received signals, while camera-based methods require a single camera and simple pre-processing (e.g. resizing) on a taken image. Although some recent studies proposed (off-the-shelf) WiFi-based methods (Wang et al. 2019; Jiang et al. 2020), multiple antennas have to be strategically deployed and the persons must be positioned in between the deployed antennas. We propose a framework that uses an easy-to-install portable radar equipped with commercial UWB chips, and raw RF signals as input without complicated pre-processing works.

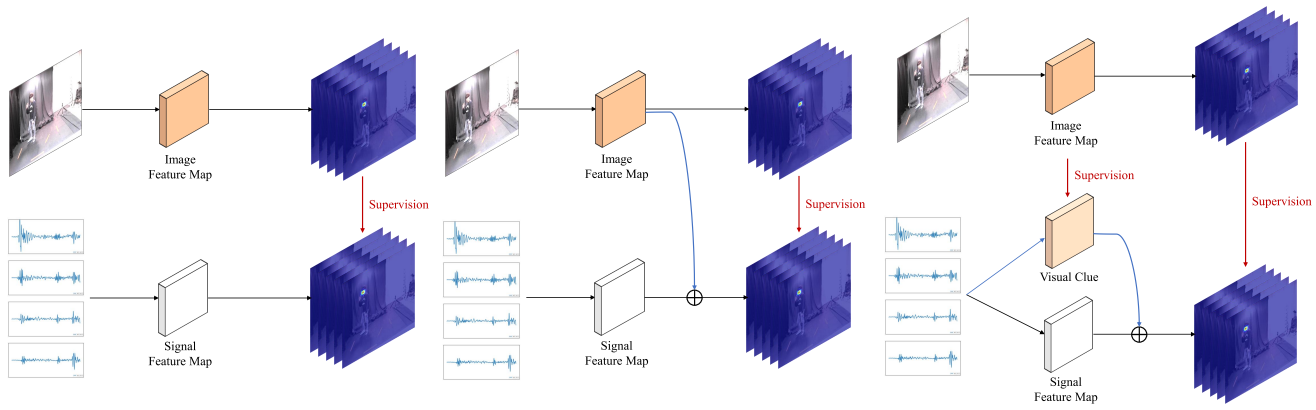
The existing RF-based methods follow a two-stage framework that is divided into a top-down and a bottom-up methods. In the top-down method, each person is first detected, and then single-person pose estimation is performed. With the bottom-up method, all possible keypoints are detected first, and then keypoints of the same person are grouped. Here lies the second challenge. Both methods require heuristic and hand-crafted post-processing such as RoI cropping (He et al. 2020), Non-Maximum Suppression (NMS) (Bodla et al. 2017), and keypoint grouping (Cao et al. 2021; Newell, Huang, and Deng 2017). These separated stage approaches make end-to-end optimization difficult. In this work, we design an RF-based Pose Estimation framework with Transformer (RPET), which operates in a fully end-to-end fashion for the first time.

The third challenge is that the RF-based methods show lower performance compared to camera-based methods. Image data used in the camera-based method is a direct representation of human body joints. Each RF signal, on the other hand, consists of waves reflected in all directions. It is very difficult to extract accurate 2D human poses from 1D RF signals. In this paper, we introduce a concept of Visual Clue (VC) which mimics an image feature map represented in a camera-based method. By learning the VC, RPET can better understand RF signal patterns and generate better RF feature representations, thus improving the accuracy of the MPPE.

The contribution of this paper is as follows.

- We designed the first fully end-to-end MPPE framework, RPET, based on RF signals. RPET is easy to deploy by using a portable radar consisting of commercial UWB chips, and uses raw RF signals without complicated pre-

*corresponding author.



(a) Traditional Cross Modal Supervision (b) Use Image Feature Map as Visual Clue (c) Cross Modal Supervision with Visual Clue

Figure 1: Cross-Modal Supervision Types for RF-based Pose Estimation: (a) is a traditional approach. (b) is a special case of adding an image feature map to RF signal feature map. (c) is our proposed model for training Visual Clue to mimic the image feature map.

processing works.

- RPET introduces the concept of Visual Clue (VC) that mimics an image feature map from the camera-based method. By training with VC, RPET can better detect individual persons and improve the accuracy of pose estimation.
- We have shown that using VC can help enhance the generalization performance of learning, especially when obstacles such as a piece of fabric or foam box (not experienced during learning) are presented, or RPET is used in different places.

Preliminary

RF-based methods follow a teacher-student design as shown in Fig. 1(a). The upper part of the figure, the teacher network, provides cross-modal supervision based on a camera-based method. It receives the image taken at the same time when receiving RF signals, and provides the pose estimation result as an annotation. The student network at the bottom of the figure receives RF signals and learns to predict the pose annotation provided by the teacher network.

Limitations of RF-based methods

A camera-based method usually encodes an image to a feature map (compressed representation) through a CNN-like backbone network. The image feature map is then upsampled (Newell, Yang, and Deng 2016; Xiao, Wu, and Wei 2018) to generate keypoint heatmaps. An image feature map has translation equivariance with an input image, and therefore direct visual information can be efficiently used for pose estimation.

On the other hand, a one-dimensional RF signals do not have a clear translation equivariance with a 2D pose estimation, so even if the RF-based method uses the same model architecture as the camera-based method, the pose estimation performance is inevitably inferior.

Feasibility of Visual Clue

As a preliminary study, we checked whether the performance of the RF-based method would be improved if the image feature map, generated by the camera-based method, was used. As shown in Fig. 1(b), the signal feature map of the student network is added to the image feature map of the teacher network. Table 1 shows that the performance can be greatly improved just by taking the image feature map as a kind of visual clue (VC). The architecture of the teacher network and student network used in this experiment will be described in detail in the Methodology section of this paper. By using the image feature map together, AP_{50} for human detection improved from 78.9 to 90.7, and pose estimation mAP improved from 65.1 to 78.1. Inspired by this result, we study to examine whether it is possible to learn VC, which mimics an image feature map, using only RF signals, and whether the learned VC can improve the performance of person detection and pose estimation.

Method	AP_{50}^{Box}	$PCKh_{50}^{Pose}$
(a) Baseline (no visual clue)	78.9	65.1
(b) With image feature map	90.7	78.1

Table 1: Preliminary Results : (a) uses a baseline model trained by a traditional cross modal supervision, (b) trains a baseline model with an image feature map from a camera-based method.

End-to-End Pose Estimation

Existing RF-based methods are designed in a two-stage framework (either top-down or bottom-up). The top-down method finds each person's bounding box in the first stage, and passes it to the next stage of single person pose estimation. RF-Pose3D (Zhao et al. 2018b) samples potential person regions on the features map as in Mask R-CNN (He

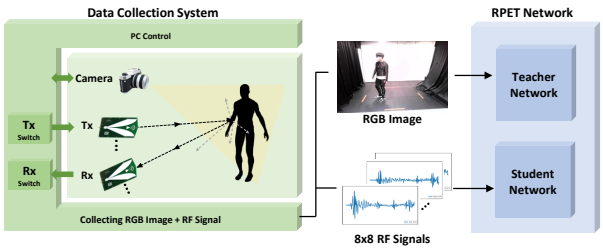


Figure 2: System Overview

et al. 2020), and needs post-processing such as hand-crafted RoI cropping and NMS. Not surprisingly, pose estimation performance is highly dependent on the performance of a person detection. The bottom-up method (Zhao et al. 2018a; Wang et al. 2019)first detects all potential keypoints in the first stage, and tries to group them into each person in the second stage. The grouping process is usually heuristic, use hand-crafted tricks with additionally learned Part Association Field (PAF) (Cao et al. 2021) to associate the detected keypoints that belong to the same person. Both two-stage methods have shown competitive performance, but are often not optimized in a fully end-to-end fashion. In this study, for the first time, we design a fully end-to-end method based on RF signals for MPPE.

System Overview

Our proposed RPET framework predicts multi-person pose estimation by taking the reflected RF signals as input. Fig. 2 shows a system overview which includes data collection and cross-modal supervision.

Data Collection

We have built a portable radar using a commercial UWB chip, Novelda NVA-6100. The radar transmits a short UWB pulse, and receives the reflected signals. Our radar operates with 8x8 antennas (8 transmitters and 8 receivers), and a single RF frame consists of 64 signals. In addition to the collection of RF data, we also use a RGB camera to obtain pose estimation annotations. We collect training data at four different locations in a room of 5m X 5m. For training data, there is no obstacle between our radar and the persons to be estimated of their poses. For testing data, we collect data (1) at different locations from the training data but in the same room, (2) in a different room, and (3) with added obstacles (e.g. a piece of fabric and a foam box) in front of the radar. For the training data collection, 12 volunteers performed daily activities, and the number of coexisting persons in a single frame varied from 1 to 4 as shown in Table 2. The total number of collected RF frames for train is 176,000.

Remove Ambient Static Information

RF signals are easily affected by ambient changes. Therefore, it is important to capture only person relevant information for MPPE, excluding any information on the surrounding objects and walls. This can considerably affect the learning efficiency and the generalization performance under various environment changes. To address this, when training a RF-based model, a moving average filter is utilized to subtract the signal's average value over a specified period time, and portions reflected by static objects can be excluded. This allows the model to only retain information that changes over a short-term interval, such as a moving person. In this paper, environment information was removed by subtracting the average of 64 consecutive frame signals from the current frame signal.

Num.	1	2	3	4	Total
N	87,000	53,000	24,000	12,000	176,000
Loc.	A	B	C	D	Total
N	34,000	33,000	49,000	60,000	176,000

Table 2: Training Dataset : we collect training data at four different locations A, B, C, D in the same room. The maximum and average number of persons in a single RF frame are 4 and 1.78, respectively.

Cross Modal Supervision with Visual Clue

As a cross-modal supervision method, we employ teacher-student network structure for label generation and training shown in Fig. 1(c). The pre-trained camera-based object detection and pose estimation models are used as a teacher network to generate ground truth for supervised learning. Cascased r-cnn (Cai and Vasconcelos 2018) and HRNet (Wang et al. 2021) are used to generate bounding box labels and pose estimation labels. As a teacher network for learning visual clue, we employ HigherHRNet (Cheng et al. 2020) to obtain image feature maps that contain human pose-specific information.

Methodology

Overall Architectures

The overall architecture consists of (1) a feature encoder, (2) a Person Detection Network (PDN), and (3) a pose estimator. The feature encoder extracts a feature map that incorporates signal and visual clue from RF signals. The PDN is built to detect the bounding box for an individual person. The pose estimator predicts body keypoints for each person. The three modules are trained in a fully end-to-end fashion without any post-processing works.

Feature Encoder A good represented feature map is important for MPPE. Our encoder refines ConvNeXt (Liu et al. 2022), a ResNet-based model primarily used in traditional computer vision tasks. We split the encoder backbone into two branches; One branch is to represent an RF signal feature map and the other is for a feature map of Visual Clue (VC). VC feature map is to mimic an image feature map obtained from the teacher backbone network. The signal and VC feature maps are then concatenated, called the integrated map. We apply some regularization methods such as stochastic depth (Huang et al. 2016), Layer Scale (Tou-

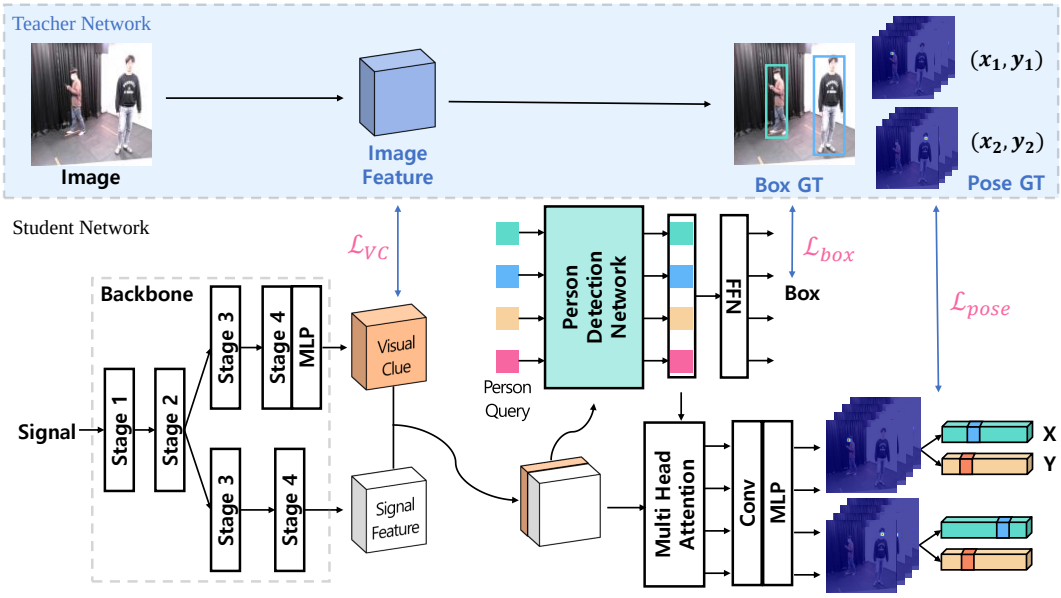


Figure 3: RPET Architecture. A top pipeline generates ground truth including bounding boxes, body keypoints, and image feature maps. A bottom pipeline consists of three modules; a Feature Encoder to generate the integrated feature map of both RF signals and Visual Clue, a Person Detection Network to detect a bounding box for each person, and a Pose Estimator to predict a body's keypoints. The three modules are trained in a fully end-to-end fashion.

vron et al. 2021) and dropblock (Ghiasi, Lin, and Le 2018) into the backbone network to enhance learning efficiency.

Person Detection Network Inspired by DETR (Carion et al. 2020), we design an object query-based person detection network to extract individual person region from the integrated feature map, including knowledge about all people represented by the feature encoder. The PDN performs a multi-head attention between the integrated feature map and the object query, and estimates the bounding box by a feed forward network (FFN).

Pose Estimator The object queries, the output of PDN, include individual person information from the observed space. Using the output of PDN, a pose estimator can perform individual pose estimation. The pose estimator computes a multi-head attention score between the queries and the integrated feature map from the encoder, and then generates as many attention heatmaps as the number of object queries. The attention heatmaps are fed into convolution layers and multi-layer perceptron (MLP), and finally outputs the prediction of x and y coordinates of each body keypoint.

Training Process

For a given RF frame which consists of 64 signals, the PDN detects a bounding box, in which a person is located in a 2D space. The PDN is trained by set prediction loss on a fixed-size bounding box coordinates prediction. The loss of the bounding box is computed by the linear combination of L1 loss and generalized IOU (Rezatofighi et al. 2019) loss between the predicted and the ground-truth box coordinates.

$$L_{box} = \lambda_{L1} L_{L1} + \lambda_{iou} L_{iou} \tag{1}$$

For pose estimation training, as demonstrated in previous works (Carion et al. 2020; Sun et al. 2021), we use bipartite matching losses for set prediction.

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum L_{box} L_(b_i, \hat{b}_{\sigma(i)}) \tag{2}$$

where b_i is a vector of ground truth box coordinates, $\hat{b}_{\sigma(i)}$ of predicted box. $\hat{\sigma}$ is the optimal assignment computed based on the Hungarian algorithm (Kuhn 2010). The ground truth and prediction are matched, based on the matching cost between the boxes utilized in the first stage. Since all pose estimations are conducted for persons within the predicted bounding box, additional pose based matching cost was not applied.

As a label for pose estimation, we use keypoint heatmaps indicating the coordinates of each body joint in a gaussian distribution and the precise x, y coordinates. Pose estimator predicts the keypoint heatmaps and compares it to the ground truth. In traditional heatmap-based pose estimation, the refinement post process of estimating the final joint coordinates based on the max value in the keypoint heatmaps is indispensable. In this paper, inspired by SimCC (Li et al. 2021b), the final coordinates of the keypoints are predicted in the form of $o_x \in [1, W_x], o_y \in [1, H_y]$ directly via FFN head, without the heuristic post-process. Kullback-Leibler(KL) divergence is adopted as the loss function in coordinate classification. Label smoothing (Szegedy et al. 2016), which is commonly employed to enhance the performance of classification models, is also applied.

$$L_{pose} = \lambda_{hm} L_{hm} + \lambda_{cc} L_{cc} \tag{3}$$

Visual clue is trained auxiliary using the L2 loss between the visual clue predicted by backbone through signals and the image feature map constructed by the feature network. Overall loss is calculated as follows.

$$\mathcal{L} = \lambda_{box} L_{box} + \lambda_{pose} L_{pose} + \lambda_{VC} L_{VC} \tag{4}$$

Experiments

Evaluation Metrics

We follow the format of keypoints introducted in MPII Human Pose (Andriluka et al. 2014), and use standard Average Precision (AP) as performance metrics. Person detection is evaluated based on the matching of bounding boxes using Intersection Over Union(IOU). Percentage of Correct Keypoints (PCK) metric is applied to evaluate pose estimation accuracy. Following the PCK, a body joint is considered to be detected correctly if the Euclidean distance between prediction and ground-truth lie within a distance of the ground truth head size (referred to as PCKh_{50}).

$$PCKh_{50} = \frac{1}{P} \sum_{p=1}^{P} \Pi(\frac{\|pred - gt\|_2^2}{headsize} \leq 0.5) \tag{5}$$

Where $\Pi(\cdot)$ outputs +1 If the conditions in parentheses are met. P is the number of persons in frames. $\frac{\|pred-gt\|_2^2}{headsize}$ represents the Euclidean distance between prediction and ground-truth, normalized by the ground-truth head size. The testing data is collected in different locations than where the training data was collected. In every test scenario, the minimum number of test data is greater than 2,000 frames.

Implementation Details

We train RPET using AdamW optimizer (Loshchilov and Hutter 2019) with a weight decay of 1e-4. There is a 10 epochs linear warm up and a cosine decaying schedule for 40 epochs. The stochastic depth rates are 0.3. Layer Scale of initial value 1 is applied. We empirically set $\lambda_{L1}, \lambda_{hm}, \lambda_{cc}, \lambda_{VC}$ as 10, λ_{iou} as 8. For Faster wall-clock time training, similar to DETR segmentation part, we train RPET for boxes only for 200 epochs, then freeze all the weights and train only the pose estimator for 30 epochs. Implementation details and hyper parameters are presented in the supplementary material.

Multi-Person Pose Estimation Results

Table 3 shows the performance results of person detection and pose estimation with and without VC. With VC, it consistently performs better than without VC, regardless of the number of people in the observed space. When the number of people increases, the prediction performance decreases as the signal reflections become more complex. Table 4, demonstrates the estimation performance of each keypoint. The results indicate that the estimation accuracy varies with body parts. This is to be expected because the amount of RF reflection depends on the size of the body part. However, it has consistently been shown that VC helps improve the prediction accuracy.

Task	VC	# of Persons				
		Total	1	2	3	4
Box		78.9	81.4	75.8	81.8	76.6
	✓	87.1(+8.2)	91.1	88.1	85.2	78.9
Pose		65.1	69.8	63.0	64.1	60.3
	✓	71.3(+6.2)	78.4	69.9	68.2	63.5

Table 3: Multi-person pose estimation results at testing locations.

VC	Hea	Nec	Sho	Elb	Wri	Hip	Kne	Ank	Total
	67.1	78.4	71.8	58.5	21.1	80.5	77.2	66.3	65.1
✓	77.7	87.3	78.5	62.9	21.9	86.3	82.9	72.8	71.3

Table 4: PCKh_{50} results of each body joint at testing locations.

Out of Distribution Generalization

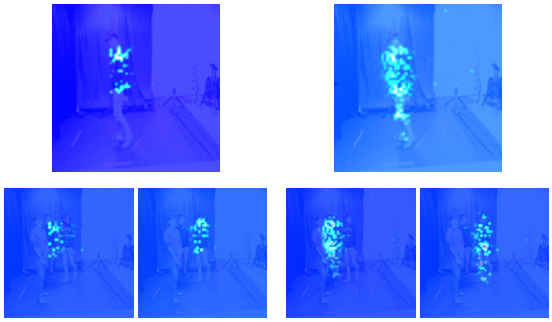
In RF-based pose estimation, when the signal acquisition location changes, the pattern of the reflected signal also changes significantly . Therefore, out of distribution generalization performance when the signal collection environment changes is essential for practical applications. We evaluate the generalization performance in a room different from the one in which the training data was collected. We call it a testing at an untrained environment. For other generalization tests, we add obstacles (such as a piece of fabric and a foam box) in front of our radar to generate invisible situations. Table 5. reports the performance of MPPE for the above scenarios. The results are slightly inferior compared to Table 3, but shows a consistent improvement in performance with VC.

Scenario	VC	Box		Pose
		mIOU	AP_{50}	PCKh_{50}
Untrained Env.		68.8	78.9	65.1
	✓	73.6	87.0	71.3
Occlusion		55.9	64.8	59.6
- Fabric	✓	64.8	76.7	61.2
Occlusion		54.4	63.1	61.9
- Foam Box	✓	60.7	74.1	66.0

Table 5: Out of distribution generalization results

Learning Efficiency with Visual Clue

Fig. 5 shows a bounding box AP_{50} converge curves in the test set when the visual clue is applied. With the aid of visual clue, we can improve the detection accuracy with less



(a) **Without** Visual Clue (b) **With** Visual Clue

Figure 4: Visualization of heatmaps for each person.

training epochs. To achieve 70 AP_{50}, 29 epochs are required when visual clue is employed, whereas 66 epochs if not. It shows a performance of 83.5 AP_{50} with visual clue and 63.1 AP_{50} without visual clue.

Fig. 4 displays the model's combined keypoint heatmaps for the same sample with or without the aid of visual clue. It shows that the heatmap of the model with VC can better cover the whole body. This better construction of the heatmaps could improve the performance of the pose estimation.

Integration Ratio of Signal Feature and Visual Clue

Our feature map is an integrated version of signal feature and visual clue. When concatenating signal feature and visual clue, we can control the degree of their influence by adjusting their channel sizes. We call it the integration ratio, and it can finally affect the performance of MPPE. To check the effect, we change the integration ratio as shown in Table 6. The results shows that when concatenating signal feature and visual clue at a ratio of 3:1, it achieves the highest performance experimentally.

Signal Feature	Visual Clue	Box		Pose
		mIOU	AP_{50}	$PCKh_{50}$
100%	0%	68.8	78.9	65.1
75%	25%	**73.6**	**87.1**	**71.3**
50%	50%	72.5	85.7	69.7
25%	75%	73.4	86.4	70.8
0%	100%	71.5	84.8	67.0

Table 6: The integration ratio between the signal feature and visual clue

Dependency on Non-Maximum Suppression

Our end-to-end framework eliminates the need for non-differentiable components such as NMS (selecting one out of many overlapping candidates). In this section, we check that adopting NMS may improve our performance additionally. Table 7 shows the performance comparison according to whether or not NMS is applied. It was confirmed that there was almost no performance improvement even if NMS was added to RPET. This shows that our model is well end-to-end optimized without NMS.

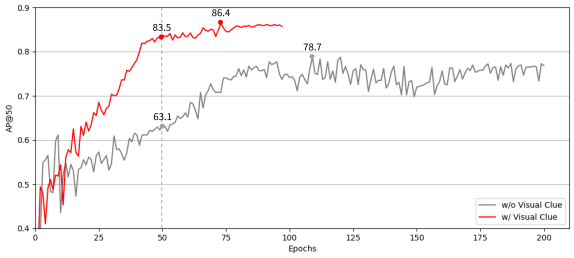


Figure 5: Convergence curves of model with or without visual clue on test set.

VC	NMS	Box		Pose
		mIOU	AP_{50}	$PCKh_{50}$
		68.8	78.9	65.1
	✓	68.1	81.0	66.8
✓		73.6	87.1	71.3
✓	✓	73.6	87.4	71.7

Table 7: Dependency on Non-Maximum Suppression

Related Work

Camera based Person Pose Estimation

Pose estimation is one of the most challenging computer vision tasks, which predicts a person's body keypoints on an image, especially in Multi-person pose estimation (MPPE). Existing methods for MPPE can be categorized into two groups: top-down and bottom-up approaches. The top-down approach (Xiao, Wu, and Wei 2018; Wang et al. 2021) detects a bounding box, which represents a person's location using object detection model (He et al. 2020). Then Estimating single person pose from cropped person instance(image or feature map). The bottom-up approach (Cao et al. 2021; Newell, Huang, and Deng 2017; Cheng et al. 2020) detects all body joints in the entire image and groups them into individual persons. Bottom up methods typically has simplified pipeline. However, heuristic grouping process relies on hand-designed components which frequently results in lower performance compared to the top-down approach.

The application of a transformer (Vaswani et al. 2017) in natural language processing and computer vision has become mainstream. ViT (Dosovitskiy et al. 2021) and Swin Transformer (Liu et al. 2021) proved that the transformer can effectively replace CNN in computer vision for various tasks. DETR (Carion et al. 2020; Zhu et al. 2021) suggests a paradigm for object detection in an end-to-end manner. By feeding a set number of learnable object queries into the

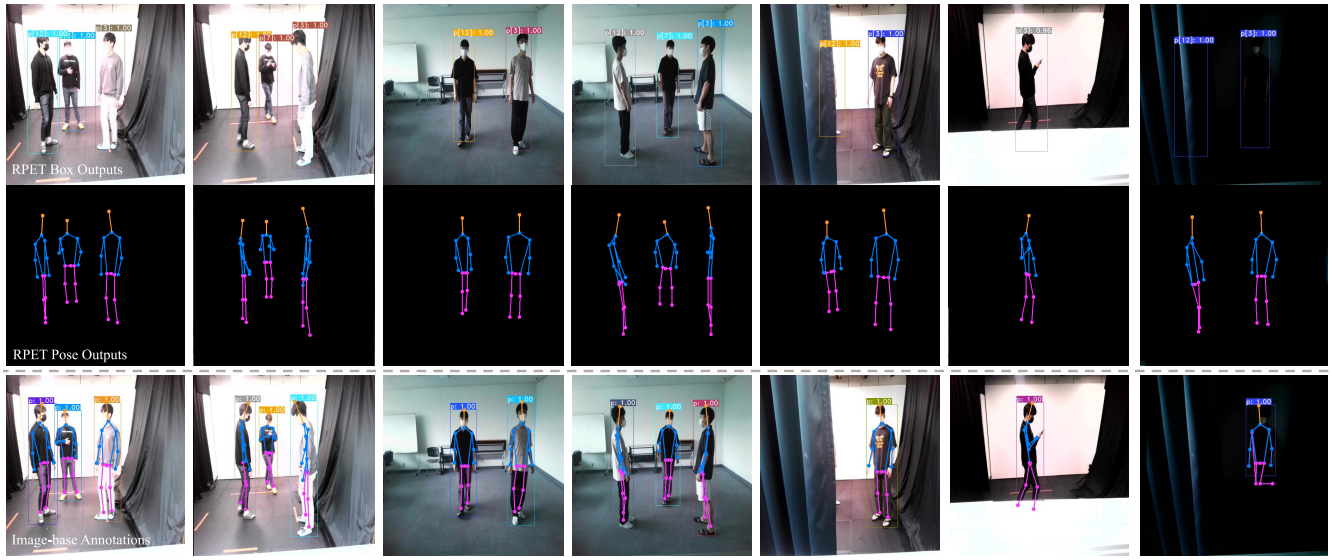


Figure 6: Qualitative Results of person detection and pose estimation on test data including different scenarios of untrained, occluded, and dark environment compared to camera based annotations. First row: Person detection results. Second row: Pose estimation results. Third row: Camera based annotations.

transformer decoder, each object query predicts the coordinates and classes of independent bounding boxes and detects multiple objects by matching the prediction with one-to-one bipartite matching on the ground-truth. TFPose (Mao et al. 2021) and PRTR (Li et al. 2021a) define MPPE as a transformer-based regression problem but still stick to the top-down structure and require hand-crafted ROI cropping components. PETR (Tian, Chen, and Shen 2019) use transformer to build a fully end-to-end framework for MPPE.

RF-based Pose Estimation

WiTrack (Adib et al. 2014) computed Time of Flight using Frequency Modulated Continuous Wave (FMCW) radar and generate the 2-dimensional(zx and yz axes) depth maps of the environment. Using a similar system, RF-Pose (Zhao et al. 2018a) applied deep learning approaches to detect body joints. However, generating depth maps require precisely built and synchronized T-shaped antenna array and a complex process. (Zheng et al. 2022) conducted pose estimations with data collected using a UWB radar. RF signals, however, require complicated pre-processing with a single target of height-azimuth dimension and multi-targets of height-azimuth and range-azimuth dimension. Person-in-Wifi (Wang et al. 2019) estimated the pose through 1-dimensional signal acquired from the multi-transceiver, as described in this work. The transmitter and receiver faced each other and estimated person's pose in between. In addition, the evaluation was conducted in the same location as the training performance, and performance dropped considerably in the untrained environment. MmMesh (Xue et al. 2021) used commercial portable millimeter-wave devices. MmMesh system achieves generating a human mesh as an enrichment of the human pose and real-time estimation. However, it can only estimate a single person. In this paper, the process of collecting radar signals through the portable radar antenna composed of multiple IR-UWB radar transceiver. Our RF-based MPPE framework achieves good generalization performance, utilizing a one-dimensional raw signal that has undergone only minimum pre-processing techniques at the average filtering level.

Conclusion

We have designed the first RF-based fully end-to-end framework with Transformer, RPET, for multi-person pose estimation (MPPE). RPET is easy to deploy with a portable radar using commercial UWB chips, and does NOT require any complicated pre-processing and post-processing works. We also introduce the concept of Visual Clue (VC) which mimics an image feature map represented in a camera-based method. Through extensive experiments, we have shown that integrating VC and signal feature can improve both accuracy and generalization performance of MPPE. In the future, we plan to take a sequence of RF frames as input to improve the pose estimation accuracy as well as extend it to individual tracking and activity recognition.

Reference

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Adib, F. M.; Kabelac, Z.; Katabi, D.; and Miller, R. 2014. 3D Tracking via Body Radio Reflections. In *NSDI*.

Zheng, Z.; Pan, J.; Ni, Z.; Shi, C.; Zhang, D.; Liu, X.; and Fang, G. 2022. Recovering Human Pose and Shape From Through-the-Wall Radar Images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Zhao, M.; Li, T.; Alsheikh, M. A.; Tian, Y.; Zhao, H.; Torralba, A.; and Katabi, D. 2018a. Through-Wall Human Pose Estimation Using Radio Signals. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7356–7365.

Zhao, M.; Tian, Y.; Zhao, H.; Alsheikh, M. A.; Li, T.; Hristov, R.; Kabelac, Z.; Katabi, D.; and Torralba, A. 2018b. RF-based 3D skeletons. *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*.

Wang, F.; Zhou, S.; Panev, S.; Han, J.; and Huang, D. 2019. Person-in-WiFi: Fine-Grained Person Perception Using WiFi. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5451–5460.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *ECCV*.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple Baselines for Human Pose Estimation and Tracking. In *ECCV*.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; Liu, W.; and Xiao, B. 2021. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 3349–3364.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2020. Mask R-CNN. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42: 386–397.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6154–6162.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 172–186.

Newell, A.; Huang, Z.; and Deng, J. 2017. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. *ArXiv*, abs/1611.05424.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *ArXiv*, abs/1706.03762.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. *ArXiv*, abs/2005.12872.

Kuhn, H. W. 2010. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; and Luo, P. 2021. Sparse R-CNN: End-to-End Object Detection with Learnable Proposals. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14449–14458.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *ArXiv*, abs/2010.04159.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; and Tu, Z. 2021a. Pose Recognition with Cascade Transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1944–1953.

Tian, Z.; Chen, H.; and Shen, C. 2019. DirectPose: Direct End-to-End Multi-Person Pose Estimation. *ArXiv*, abs/1911.07451.

Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; and Wang, Z. 2021. TFPose: Direct Human Pose Estimation with Transformers. *ArXiv*, abs/2103.15320.

Liu, Z.; Mao, H.; Wu, C.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Huang, G.; Sun, Y.; Liu, Z.; Sedra, D.; and Weinberger, K. Q. 2016. Deep Networks with Stochastic Depth. In *ECCV*.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and J'egou, H. 2021. Going deeper with Image Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 32–42.

Ghiasi, G.; Lin, T.-Y.; and Le, Q. V. 2018. DropBlock: A regularization method for convolutional networks. In *NeurIPS*.

Rezatofighi, S. H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666.

Li, Y.; Yang, S.; Liu, P.; Zhang, S.; Wang, Y.; Wang, Z.; Yang, W.; and Xia, S. 2021b. SimCC: a Simple Coordinate Classification Perspective for Human Pose Estimation.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception Architecture for Computer Vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826.

Andriluka, M.; Pishchulin, L.; Gehler, P. V.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 3686–3693.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS — Improving Object Detection with One Line of Code. *2017 IEEE International Conference on Computer Vision (ICCV)*, 5562–5570.

Yue, S.; Yang, Y.; Wang, H.; Rahul, H.; and Katabi, D. 2020. BodyCompass: Monitoring Sleep Posture with Wireless Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4: 66:1–66:25.

Jiang, W.; Xue, H.; Miao, C.; Wang, S.; Lin, S.; Tian, C.; Murali, S.; Hu, H.; Sun, Z.; and Su, L. 2020. Towards 3D human pose construction using WiFi. In *Proceedings of the*

26th Annual International Conference on Mobile Computing and Networking, 1–14.

Xue, H.; Ju, Y.; Miao, C.; Wang, Y.; Wang, S.; Zhang, A.; and Su, L. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 269–282.

Implementation Details

config	RPET
optimizer	AdmaW (Loshchilov and Hutter 2019)
base learning rate	2e-4
weight decay	1e-4
optimizer momentum	β_1, β_2=0.9,0.999
batch size	128
training epochs	200(box) + 30(pose)
learning rate schedule	cosine decay
warmup epochs	10
warmup schedule	linear
label smoothing (Szegedy et al. 2016)	0.1
stochastic depth (Huang et al. 2016)	0.3
layer scale (Touvron et al. 2021)	1.0
dropblock (Ghiasi, Lin, and Le 2018)	0.2
$\lambda_{L1}, \lambda_{iou}, \lambda_{hm}, \lambda_{cc}, \lambda_{VC}$	10, 8, 10, 10, 10
PDN # of layers	6
PDN # of heads	8
PDN # of object queries	15

Table 8: Implementation details and hyper parameters