Neural Interactive Keypoint Detection

Jie Yang[1,2*], Ailing Zeng[1†], Feng Li[1], Shilong Liu[1], Ruimao Zhang[2†], Lei Zhang[1]
[1]International Digital Economy Academy
[2]School of Data Science, Shenzhen Research Institute of Big Data,
The Chinese University of Hong Kong, Shenzhen
{jieyang5@link,zhangruimao@}cuhk.edu.cn
{zengailing,lifeng,liushilong,leizhang}@idea.edu.cn
https://github.com/IDEA-Research/Click-Pose

(a) In-domain annotation (b) Out-of-domain annotation (c) 4 extra keypoints annotation

Figure 1: We demonstrate the effects of *Click-Pose* in three keypoint annotation scenarios. For scenarios (a) and (b), the left figures show the model-only's initial keypoint localization, followed by the corrected keypoints (red points) obtained through user clicks. The right figures display the final results obtained by *Click-Pose* after automatically refining other keypoints and corresponding human boxes. For scenario (c), the left figure illustrates the original task of detecting 17 keypoints, while the right figure shows the adaptability of *Click-Pose* in detecting additional 4 keypoints.

Abstract

This work proposes an end-to-end neural interactive keypoint detection framework named Click-Pose, which can significantly reduce more than 10 times labeling costs of 2D keypoint annotation compared with manual-only annotation. Click-Pose explores how user feedback can cooperate with a neural keypoint detector to correct the predicted keypoints in an interactive way for a faster and more effective annotation process. Specifically, we design the pose error modeling strategy that inputs the ground truth pose combined with four typical pose errors into the decoder and trains the model to reconstruct the correct poses, which enhances the self-correction ability of the model. Then, we attach an interactive human-feedback loop that allows receiving users' clicks to correct one or several predicted keypoints and iteratively utilizes the decoder to update all other keypoints with a minimum number of clicks (NoC) for efficient annotation. We validate Click-Pose in in-domain, out-of-domain scenes, and a new task of keypoint adaptation. For annotation, Click-Pose only needs 1.97 and 6.45 NoC@95 (at precision 95%) on COCO and Human-Art, reducing 31.4% and 36.3% efforts than the SOTA model (ViTPose) with manual correction, respectively. Besides, without user clicks, Click-Pose surpasses the previous end-to-end model by 1.4 AP on COCO and 3.0 AP on Human-Art.

1. Introduction

Multi-person keypoint detection aims to localize 2D coordinates of keypoints for each person in images, as in Fig. 1. It has garnered significant attention in research and industry, particularly in sports, entertainment, and surveillance applications. The development of deep models for various applications heavily depends on a large volume of training data with labels (e.g., COCO [12, 23]). As the amount of data increases, the manual annotations of dense human keypoints are quite time-consuming, labor-intensive, and cost-prohibitive. As demonstrated in Fig. 2, annotating a single person with 17 keypoints would take about 230 seconds. For a dataset of $50K$ images with an average of four people per image, this process would require 532 hours. Additionally, there may exist omissions, localization deviation, and mislabeling in the manual annotation process.

To reduce the manual effort, an intuitive annotation process can use a state-of-the-art (SOTA) model [42] to obtain

*Work done during an internship at IDEA.
†Corresponding author.

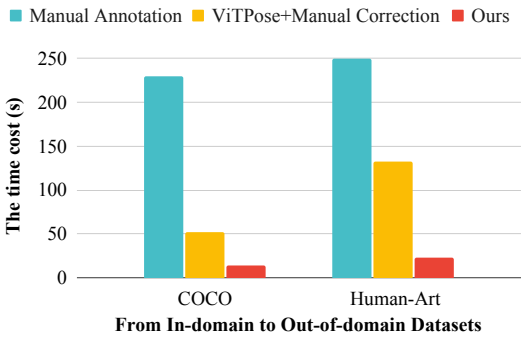


Figure 2: Comparison of the average time cost of keypoint annotation per person using three strategies on two datasets. Our proposed *Click-Pose* is more than **10** times faster than manual annotation. Importantly, it significantly alleviates model bias in out-of-domain annotation (e.g., on Human-Art), reducing the time required by **83**% compared to state-of-the-art model annotation with manual correction.

a preliminary model-annotation result and then mannually correct all wrong keypoints. However, this strategy heavily relies on the performance of the model to reduce manual effort, which leads to the following problems: 1) **model bias**. As shown in Fig. 2, for in-domain data, the SOTA model (e.g., ViTPose-H) can accelerate the annotation process by about four times due to its high prediction accuracy. However, its performance may be suboptimal when labeling an out-of-distribution (OOD) dataset (e.g., Human-Art [13]) or when dealing with new keypoints [40] that have not been defined. In such cases, more manual efforts will be required. 2) **performance bottleneck**. The performance of existing SOTA models is generally hard to be further improved, which makes it challenging to further reduce manual effort. Noticing that there exist inherent problems in both the manual-only annotation and the model with manual correction strategies, the following questions naturally arise: *how can we integrate manual correction with model predictions in an interactive manner to enable faster, more accurate, and more versatile keypoint annotation with minimal user correction?*

To address this issue, we define a novel task called interactive keypoint detection. It aims to effectively maximize benefits of the model to minimize manual effort, and mitigate unfriendly consequences of model failures in out-of-distribution and new-task annotations that increase the need for manual intervention. Accordingly, we present the first neural interactive keypoint detection framework, *Click-Pose*, as a baseline for further research. It allows a user to direct correct the positions of one or multiple keypoints and incorporate this feedback to refine other keypoints in Fig. 1.

Specifically, we build *Click-Pose* upon the end-to-end SOTA model ED-Pose [43]. This model incorporates a keypoint-to-keypoint refinement scheme through a regression head and updates keypoints layer-by-layer in the decoder, which allows receiving user-corrected positions at the decoder instead of the input image. However, we empirically find that the decoder in ED-Pose is extremely susceptible to variations in input keypoint positions. Even a minor deviation can result in a significant deterioration in performance. To tackle this limitation, we introduce two unique technical contributions to its decoder. The first is the pose error modeling that builds a reconstruction task to enhance the robustness of the decoder and learn to refine wrong keypoints by leveraging the correct keypoints as a reference. The second is the interactive human-feedback loop, which allows receiving users' clicks to correct one or several predicted keypoints and iteratively utilizes the decoder to update all other keypoints with minimal manual corrections for efficient annotation.

Click-Pose incorporates the above two essential designs into the training process, which improves $+1.4$ AP on COCO `val` and $+3.0$ AP on HumanArt `val` compared with the baseline model ED-Pose, achieving state-of-the-art performance for end-to-end keypoint detection. More importantly, as shown in Fig. 1, *Click-Pose* shows its advantages in various annotation scenarios, *i.e.,* in-domain, out-of-domain scenes, and a new task of keypoint adaptation. Specifically, *Click-Pose* only needs 1.97 and 6.45 NoC@95 (the average number of user clicks needed to annotate one person to achieve a precision of 95%) on COCO and Human-Art, reducing 31.4% and 36.3% efforts than the SOTA model with manual correction, respectively. Moreover, *Click-Pose* significantly reduces the average time cost of single-person annotation, achieving over $5\times$ speedup compared to the SOTA model ViTPose with manual correction and more than a $10\times$ speedup compared to manual-only annotation, especially in out-of-domain scenarios.

Our contributions are: (1) We define a novel task called interactive keypoint detection to pursue high-precision and low-cost annotation, and present the first framework to address this task, namely *Click-Pose*. (2) We incorporate the pose error modeling and interactive human-feedback loop into the training of *Click-Pose*, leading to a state-of-the-art performance for end-to-end keypoint detection. (3) We provide a new metric (NoC) and extensively validate the effectiveness and efficiency of *Click-Pose* in different annotation scenes. We hope this work could inspire further research in related fields.

2. Related work

2.1. Multi-Person Pose Estimation

Existing pose estimation models can be generally divided into two-stage methods and one-stage methods. For two-stage methods, there are top-down (TD) and bottom-up (BU) strategies. Top-down methods [22, 27, 35, 38, 42] have achieved high performance by first detecting each per-

son in the image with an object detector and then conducting the single-person pose estimation with the proposed model. However, these methods are limited by their inability to handle missing person detections and their high costs for crowd scenes. In contrast, bottom-up methods [2, 5, 9, 26, 29] have demonstrated greater efficiency by first estimating keypoints and then grouping them into individual human poses. Recently, the advent of end-to-end object detectors DETR [3] has led to the development of one-stage pose estimators, like PETR [32] and ED-Pose [43]. ED-Pose is particularly notable in that it approaches this task as two explicit box detection processes, leading to superior performance and efficiency trade-offs. Additionally, some refinement models [14, 28, 45, 46] also focus on pose correction. They take both the original image and an estimated pose as inputs to refine a more accurate pose. However, despite great efforts to achieve state-of-the-art models, such as ViTPose [42] with ViT-Huge backbone [7], and other models [14, 28] that specialize in pose refinement, they still require manual correction to satisfy the precision requirement of annotation, where even more manual effort is required to compensate for the performance drop in out-of-distribution annotation scenarios. In contrast, we attempt to address interactive keypoint annotation with minimal manual efforts using a fully end-to-end framework.

2.2. Human-in-the-Loop

Annotation is a typical application scenario for Human-in-the-Loop (HITL) techniques, which aims to improve prediction models' accuracy while minimizing costs by leveraging human knowledge and experience. Existing works mainly focus on two directions: (i) data processing via human feedback. For instance, one approach, known as *active learning* [1, 16], seeks to minimize manual annotation effort on a large dataset while maximizing the model's performance [8, 33, 39]; In specific, prior HITL methods for pose estimation [8, 10, 24] have involved actively selecting and labeling informative images to facilitate effective learning. (ii) interventional model training and inference via human feedback. For instance, the interactive image segmentation task is to extract an accurate target mask with minimal user interaction [4, 25, 34, 41]. This is a popular research area over the past years. Existing deep learning-based approaches usually input both the image and user annotations in the model training and testing stages or conduct various inference-time optimization schemes [11, 19], which suffer from high computation costs and slow speeds for each annotation. A prior study [15] introduces an interactive image segmentation pipeline designed for heatmap-based interactive keypoint annotation in X-ray images. Additionally, researcher [6] has delved into the realm of inter- and extrapolated annotations across frames. Remarkably, *no work has yet explored how to enable effective interaction between deep models and human feedback to improve end-to-end multi-person keypoint annotation accuracy with fewer costs and manual efforts*. In this work, inspired by the concept of HITL, we first investigate how to combine human feedback with a deep model in an interactive manner for human body keypoint annotation.

3. Methodology

3.1. Motivation

Interactive Keypoint Detection. Interactive keypoint detection aims to obtain accurate keypoint annotations with minimal user interactions. For example, if a network predicts an incorrect pose, such as a flipped pose, the user may only need to correct one keypoint by clicking on it. Subsequently, the network can use this human feedback to further refine the remaining keypoint positions and determine the correct pose. To address this task, the network should incorporate a pose-to-pose refinement scheme that can receive modified keypoint positions from the user and output further refined positions.

Preliminary Study of ED-Pose [43]. ED-Pose addresses the task of keypoint detection by explicitly reformulating 4D keypoint boxes as queries and progressively refining them layer by layer in the decoder through a regression head. It achieves SOTA performance compared with existing end-to-end models and improves the inference speed. When considering only the 2D coordinate, ED-Pose can be seen as providing a keypoint-to-keypoint refinement scheme, thus conceptually satisfying the aforementioned architectural requirement for interactive keypoint detection. As shown in Fig. 3-(a), it consists of an Encoder, a Human decoder, and a Human-to-Keypoint decoder. Specifically, it extracts image features using a backbone and passes them through the Encoder with positional embedding to obtain refined image features $\mathbf{F}$. In the Human decoder, ED-Pose leverages human queries $\mathbf{Q}_H$ to search for human objects, where $\mathbf{Q}_H$ contains position queries $\mathbf{Q}_H^p$ (*i.e.,* human box positions) and content queries $\mathbf{Q}_H^c$ (*i.e.,* human content embedding). Then, it utilizes the updated human queries $\mathbf{Q}_H'$ to initialize keypoint queries $\mathbf{Q}_K$, where $\mathbf{Q}_K$ also includes position queries $\mathbf{Q}_K^p$ (*i.e.,* keypoint positions) and content queries $\mathbf{Q}_K^c$ (*i.e.,* keypoint content embedding). Finally, it attaches the Human-to-keypoint decoder to refine the human box and keypoints of each person to $\mathbf{Q}_H''$ and $\mathbf{Q}_K'$.

Non-interactive Issue in ED-Pose. As mentioned above, the Human-to-Keypoint decoder of ED-Pose can be viewed as a keypoint-to-keypoint refinement process. It is natural to consider whether human feedback (e.g., a corrected keypoint) can be directly incorporated in the decoder without any further modification. However, extensive preliminary experiments have shown that the decoder is highly sensitive to the input keypoint position query $\mathbf{Q}_K^p$ and is

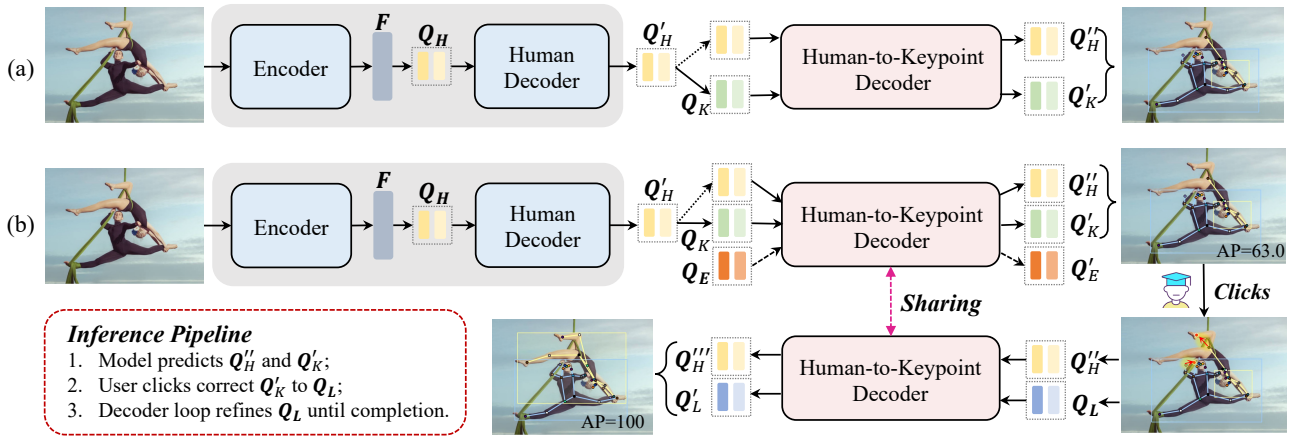


Figure 3: Comparison of (a) ED-Pose [43] with (b) the proposed *Click-Pose*. *Click-Pose* follows the same architecture as ED-Pose while introducing two key techniques to the Human-to-Keypoint decoder. Firstly, we introduce a ***training-only*** strategy, namely **Pose Error Modeling**. It builds a reconstruction task to self-correct error-keypoint queries $\mathbf{Q}_E$ to $\mathbf{Q}'_E$, which enhances the robustness of the model and learns to refine wrong keypoints by leveraging correct keypoints as a reference. Secondly, we attach an **Interactive Human-Feedback Loop** to allow the user to correct one or several keypoints positions in $\mathbf{Q}'_K$ and generate the modified keypoint queries $\mathbf{Q}_L$. Then the Human-to-Keypoint decoder could take the predicted boxes $\mathbf{Q}''_H$ and $\mathbf{Q}_L$ as input again and further refine human boxes and all keypoints to $\mathbf{Q}'''_H$ and $\mathbf{Q}'_L$ based on user corrections.

unable to effectively utilize human feedback. For example, during the inference, we randomly add a small disturbance $(\Delta x, \Delta y)$ to each keypoint coordinate (x, y) in $\mathbf{Q}^p_K$. We ensure that $|\Delta x| < \omega_x$ and $|\Delta y| < \omega_y$, where $\omega_x, \omega_y \in (0, 0.1)$. This operation results in a sharp drop in accuracy from 71.6AP to 11.8AP. There are two main reasons for this sensitivity: Firstly, the Human-to-Keypoint decoder effectively learns the contextual information of each keypoint, which leads to a strong coupling between the position query $\mathbf{Q}^p_K$ and the content query $\mathbf{Q}^c_K$. Once $\mathbf{Q}^p_K$ changes, this misalignment can cause the final results to drop off. Secondly, the Human-to-Keypoint decoder also creates a contextual coupling relationship among different keypoints, meaning that adjusting the input position of one keypoint may severely harm the update of the others.

3.2. The Overview of Click-Pose

Introduction to *Click-Pose*. As illustrated in Fig. 3-(b), *Click-Pose* adopts the same modules as ED-Pose to obtain the person box $\mathbf{Q}''_H$ and keypoints $\mathbf{Q}'_K$ from an input image. Furthermore, *Click-Pose* introduces two key techniques to the Human-to-Keypoint Decoder, which makes the model interactive and robust. Firstly, we additionally introduce error-keypoint queries $\mathbf{Q}_E$ in the training stage, which includes four typical pose errors defined by [30]. We feed them into the Human-to-Keypoint decoder to reconstruct the accurate pose $\mathbf{Q}'_E$. This operation enhances the self-correction ability of the Human-to-Keypoint decoder. We call it ***Pose Error Modeling*** (see Sec. 3.3). Secondly, we introduce the user interaction in an attached Human-to-Keypoint decoder via proposed ***Interactive Human-Feedback Loop*** (see Sec. 3.4). In this process, the user can correct one or several keypoints positions in $\mathbf{Q}'_K$ and generate the modified keypoint queries $\mathbf{Q}_L$. Then, the Human-to-Keypoint decoder could take $\mathbf{Q}''_H$ and $\mathbf{Q}_L$ as input iteratively and further refine human boxes to $\mathbf{Q}'''_H$ and all keypoints to $\mathbf{Q}'_L$ based on the modified keypoints. This process can further improve the self-correction ability of the model during the training and successfully allow the user clicks to be integrated into the inference phase.

Training Optimization Processes of *Click-Pose*. *Click-Pose* is an end-to-end trainable framework that extends the ED-Pose training process (the loss as $\mathcal{L}_g$). Firstly, *Pose Error Modeling* uses ground-truth keypoints to generate erroneous poses and creates a pose reconstruction task, introducing the loss as $\mathcal{L}_r$. Secondly, *Interactive Human-Feedback Loop* uses the ground-truth keypoints to simulate user clicks for correcting a few wrong keypoints in model predictions and loop decoder to refine other wrong keypoints and the corresponding human boxes, introducing the loss as $\mathcal{L}_l$. Finally, the overall training pipeline of *Click-Pose* can be written as follows,

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_r + \mathcal{L}_l, \tag{1}$$

where we employ a set-based Hungarian matching to ensure a unique prediction for each ground-truth pose [3, 32, 43]. Following ED-Pose, $\mathcal{L}_g$ and $\mathcal{L}_l$ include human classification loss, human box regression loss, and human pose regression loss. $\mathcal{L}_r$ is the $L1$ loss for pose reconstruction.

Inference Pipeline of *Click-Pose*. Given an image,

Click-Pose first performs an end-to-end inference to obtain all human boxes $\mathbf{Q}''_H$ and keypoint locations $\mathbf{Q}'_K$ without any troublesome post-processing. Notably, $\mathbf{Q}_E$ is not required in inference. In annotation scenes, the user can directly correct the wrong keypoints in model predictions $\mathbf{Q}'_K$ to $\mathbf{Q}_L$ and loop decoder to refine human boxes $\mathbf{Q}'''_H$ and all keypoints $\mathbf{Q}_L$ until the annotation is completed.

3.3. Pose Error Modeling

Pose Error Modeling aims to enhance the robustness of the Human-to-Keypoint decoder via a reconstruction scheme [20]. To achieve this, we introduce error-keypoint queries $\mathbf{Q}_E$ by adding four typical error types into ground-truth keypoints, and then we feed $\mathbf{Q}_E$ into the Human-to-Keypoint decoder to reconstruct the accurate poses $\mathbf{Q}'_E$.

In specific, $\mathbf{Q}_E$ consists of position queries $\mathbf{Q}^p_E$ and content queries $\mathbf{Q}^c_E$, where the former can be initialized by the 2D coordinates of keypoints and the latter can be initialized by the keypoint label embedding via a learnable codebook $\mathbf{B} \in \mathbb{R}^{K\times C}$. K is the number of defined keypoints, and C is the channel dimension. Then, we simulate four typical error types of the keypoint, *i.e., jitter*, *miss*, *swap* and *inversion* defined by [28, 30], and add them into the ground-truth keypoints for the initialization of $\mathbf{Q}^p_E$ and $\mathbf{Q}^c_E$.

For localization issues, *i.e., jitter*, *miss*, and *swap*, we perturb the ground-truth keypoints with different magnitudes of position disturbance to initial $\mathbf{Q}^p_E$. Specifically, we add a random disturbance $(\Delta x, \Delta y)$ to the (x, y) of the keypoint and make sure that $|\Delta x| < \frac{\lambda_x w}{2}$ and $|\Delta y| < \frac{\lambda_y h}{2}$, where $\lambda_x, \lambda_y \in (0, 1)$. Such disturbance constrains the keypoints with pose errors to remain within the bounding box. In addition, we use $\mathbf{B}$ directly to embed ground-truth keypoint labels to initialize the $\mathbf{Q}^c_E$. Moreover, *inversion* is a complex error that involves mislabeling and mislocating body parts within the same person (e.g., confusing the left and right elbow). As perturbing positions of ground-truth keypoints to initialize $\mathbf{Q}^p_E$, we have a hyper-parameter α (e.g., 0.4) to randomly flip the labels of the left and right body parts for the initialization of $\mathbf{Q}^c_E$. Such keypoint flipping introduces a misalignment between $\mathbf{Q}^p_E$ and $\mathbf{Q}^c_E$, which compels the model to recognize the interdependence between the position and label of the body parts. At last, we preserve a subset of ground-truth keypoints in $\mathbf{Q}_E$. This enables the model to learn how to leverage the correct ground-truth keypoints as a reference to refine wrong keypoints.

3.4. Interactive Human-Feedback Loop

Interactive Human-Feedback Loop aims to interact with user clicks, minimize manual corrections and enable efficient annotation. It allows the model to receive user clicks for correcting a few predicted keypoints and iteratively utilize the proposed decoder to update all other keypoints and human boxes.

Initialization of Modified Queries $\mathbf{Q}_L$. Given the predicted keypoints $\mathbf{Q}'_K$, which contains position queries $\mathbf{Q}^{q\prime}_K$ and content queries $\mathbf{Q}^{c\prime}_K$, the user can click the one or several keypoints in $\mathbf{Q}^{q\prime}_K$ to obtain the modified position queries $\mathbf{Q}^p_L$ of $\mathbf{Q}_L$, where $\mathbf{Q}^p_L$ only have a few keypoints corrected by the user (e.g., 1 click). Since the modified position queries $\mathbf{Q}^p_L$ and the originally predicted content queries $\mathbf{Q}^{c\prime}_K$ are misaligned, we initialize the modified content queries $\mathbf{Q}^c_L$ through label embedding using the codebook $\mathbf{B}$, which is shared with the pose error modeling process.

Training and Inference Strategies. For training, we employ Hungarian matching to obtain the predicted poses that are matched with ground-truth poses in an image. Then, we can directly modify the corresponding predicted keypoints $\mathbf{Q}^{q\prime}_K$ using the ground-truth to simulate the user click operation and obtain the modified queries $\mathbf{Q}_L$. We loop decoder to refine human boxes $\mathbf{Q}''_H$ and all keypoints $\mathbf{Q}_L$ to $\mathbf{Q}'''_H$ and $\mathbf{Q}_L\prime$, which are supervised by ground-truth boxes and keypoints. For inference, to obtain quantitative results, we evaluate the effectiveness and efficiency of *Click-Pose* for annotation on existing datasets in a manner similar to the training procedure. In real annotation scenarios, *Click-Pose* enables users to provide direct feedback to complete annotation with minimal effort.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our methods on four benchmarks: COCO [23], Human-Art [13], OCHuman [47] and CrowdPose [21]. COCO consists of about 250K person instances with 17 keypoints, and provides diverse human poses in natural scenarios. On the other hand, Human-Art comprises 123K person instances with 21 keypoints, of which 17 are the same as COCO. It provides rich human poses in out-of-distribution artistic scenes. OCHuman has 8110 human pose instances that have occlusions with the maxIOU$\geq$0.5, where 32% instances are more challenging with the maxIOU$\geq$0.75. CrowdPose provides 80000 human poses with 14 labeled keypoints in the crowded scenes.

Evaluation Metrics. Inspired by interactive segmentation [25, 34], we introduce a new metric called the Number of Clicks (NoC), which measures the average number of clicks needed to annotate one person to achieve a specific target average precision (AP). We set the target AP to 85%, 90%, and 95%, denoting the corresponding measures as NoC@85, NoC@90, and NoC@95, respectively. The average NoC is calculated over images that contain person instances in the COCO `val` set or Human-Art `val` set for evaluation. Moreover, we report the overall AP when restricting the number of clicks per person, such as C1 and C3, which aims to evaluate the performance that different methods can achieve with the same human effort.

Time cost (s)	Manual-only	ViTPose+C	Ours
COCO	230±56	52±10	**14±5**
Human-Art	250±55	132±41	**23±8**

Table 1: Comparisons of **the average and standard deviation time cost** required for single-person annotation by three annotation strategies: manual-only, SOTA model with manual correction, and our *Click-Pose*.

Method	Backbone	NoC@85 ↓	NoC@90 ↓	NoC@95 ↓
COCO `val`				
ViTPose	ViT-Huge	1.46	2.15	2.87
Click-Pose	ResNet-50	**0.95**	**1.48**	**1.97**
Human-Art `val`				
ViTPose	ViT-Huge	9.12	9.79	10.13
Click-Pose	ResNet-50	**4.82**	**5.81**	**6.45**

Table 2: Comparisons of **Number of Clicks (NoC)** metrics for interactive keypoint detection.

Implementation Details. Following [3, 43], the training images are augmented by random cropping, flipping, and resizing with the shorter sides in $[480, 800]$ and the longer sides less or equal to 1333. The number of queries $\mathbf{Q}_K$ is set to 50. We use the AdamW optimizer with a weight decay of 1×10^{-4}. Our model is trained on Nvidia A100 GPUs with a batch size of 16 for 40 epochs on COCO. The initial learning rate is 1×10^{-4} and is reduced by a factor of 0.1 at the 38th epoch on COCO. The channel dimension C is set to 256. The testing images are resized to have shorter sides of 800 and longer sides less than or equal to 1333. All compared DETR-based models use the ResNet-50 backbone.

4.2. Annotation Comparisons

We investigate the advantages of *Click-Pose* in different annotation scenes, *i.e.,* in-domain natural scene (COCO), out-of-domain artificial scene (Human-Art), and crowded scenes (OCHuman and CrowdPose). The compared pose estimators comprehensively include top-down (TP), bottom-up (BU), and one-stage (OS) models.

Time Cost Comparisons. In Tab. 1, we compared *Click-Pose* with two other annotation schemes, one using manual-only annotation, and the other using ViTPose [42] to detect initial predictions, followed by manual correction for incorrect keypoints. We conduct a study where ten users annotate the same ten images (which proved challenging for direct prediction via various methods) using different strategies, and we calculate the average and variance time it took to annotate a single person. The results show *Click-Pose* significantly reduces the annotation time cost, especially in the out-of-domain annotation scene, achieving a speedup of 10 times compared to manual-only annotation and 5 times compared to the SOTA model with manual correction.

NoC Metric Comparisons. Tab. 2 shows the performance of *Click-Pose* and ViTPose in terms of the NoC metric. Our results demonstrate that *Click-Pose* with a much smaller backbone can require fewer human corrections to achieve different AP requirements compared to ViTPose, reducing manual effort by 31.4% and 36.3% when the target AP is set to 95 (NoC@95) for COCO and Human-Art, respectively. Non-interactive deep models tend to suffer from model bias and fail on OOD annotation scenes, while *Click-Pose* can significantly mitigate this problem.

4.3. In-domain Keypoint Detection

We verify the effectiveness of *Click-Pose* in comparison to other state-of-the-art methods in the in-domain scene in Tab. 3 and 4, where we train our models on COCO `train` set and validate them on COCO `val` set.

Comparison with *Model+Manual Correction* methods: We simulate manual correction on the output results by replacing worse keypoints with ground-truth. The results show that *Click-Pose* can achieve better performance compared to ED-Pose, Poseur and ViTPose, with the same amount of human effort. For instance, when modifying 4 incorrect keypoints per person, *Click-Pose* achieves 96.4 AP, which is 8.1 AP higher than ViTPose.

Comparison with *Model-Only* methods: *Click-Pose*-C0, which does not require user corrections, achieves state-of-the-art results using the same ResNet-50 backbone in a fully end-to-end manner. This remarkable performance is attributed to the effective training facilitated by pose error modeling. Notably, *Click-Pose*-C0 outperforms ED-Pose by 1.4 AP with a faster inference time.

4.4. Out-of-domain Keypoint Detection

To demonstrate the generalization ability of *Click-Pose*, we further evaluate it in the OOD scene, where we train our models on COCO `train` set and validate them on Human-Art `val` set (only 17 keypoints here) in Tab. 5 and 6.

Comparison with *Model+Manual Correction* methods: In Tab. 5 and 6, *Click-Pose* demonstrates robust performance in such out-of-domain scenario, outperforming ViTPose by 29.5 AP, Poseur by 35.2 AP and ED-Pose by 18.3 AP when clicking 3 keypoints per-person. Furthermore, *Click-Pose*'s performance remains consistent across other settings as well.

Comparison with *Model-Only* methods: *Click-Pose*-C0 outperforms all two-stage or one-stage approaches, significantly surpassing the SOTA ViTPose model by 11.8 AP. It also achieves an improvement of 3.0 AP over ED-Pose.

4.5. Crowded Scene Keypoint Detection

Tab. 7 and 8 investigates the effectiveness of *Click-Pose* in the crowded scene. Here, we compare our *Click-Pose* to the baseline model ED-Pose with the same ResNet-50 backbone. Specifically, *Click-Pose*-C0 outperforms ED-

Method	Backbone	AP↑	AP_{50}↑	AP_{75}↑	AP_M↑	AP_L↑	Time [ms]↓
Model-Only							
ViTPose† [42] (TP)	ViT-Huge	79.1	91.7	85.7	71.9	82.0	45+286
HRNet† [36] (TP)	HRNet-w32	74.4	90.5	81.9	70.8	81.0	45+112
HrHRNet† [5] (BU)	HRNet-w32	67.1	86.2	73.0	61.5	76.1	322
PETR [32] (OS)	ResNet-50	68.8	87.5	76.3	62.7	77.7	105
ED-Pose [43] (OS)	ResNet-50	71.6	89.6	78.1	65.9	79.8	51
Click-Pose-C0 (OS)	ResNet-50	73.0↑1.4	90.4	80.0	68.1	80.5	**48**↓3
Model+Manual Correction							
ViTPose-C1	ViT-Huge	82.3	90.8	86.6	78.8	87.9	-
ViTPose-C2	ViT-Huge	85.3	91.9	89.5	83.6	89.4	-
ViTPose-C3	ViT-Huge	86.7	93.8	90.3	86.7	89.4	-
ViTPose-C4	ViT-Huge	88.3	95.2	92.4	90.9	89.5	-
Neural Interactive							
Click-Pose-C1	ResNet-50	83.2 (+1.8)	96.5 (+3.4)	89.7 (+2.3)	80.1 (+2.8)	87.9 (+0.2)	-
Click-Pose-C2	ResNet-50	90.3 (+2.7)	97.8 (+3.1)	95.2 (+4.1)	88.1 (+3.1)	93.9 (+1.9)	-
Click-Pose-C3	ResNet-50	94.1 (+3.4)	98.9 (+3.4)	96.6 (+3.8)	92.6 (+3.5)	96.5 (+2.8)	-
Click-Pose-C4	ResNet-50	**96.4** (+3.9)	**99.0** (+3.3)	**97.9** (+4.3)	**95.3** (+3.8)	**97.8** (+3.3)	-

Table 3: **Comparison with representative SOTAs** on COCO `val` set. C1-C4 limits the number of clicks on a single person. *Click-Pose*-C0 is a fully end-to-end framework without user clicks. The red arrow indicates its improvement over ED-Pose [43]. The number in parentheses is the interactive model improvement via the loop refinement (ignoring the manual improvement). † denotes the flipping test. The inference time of all model-only methods is tested on an A100, except for the detector of the top-down methods, which is referred from the MMdetection (*i.e.,* 45ms).

Method	C0↑	C1↑	C2↑	C3↑	C4↑	NoC@95↓
Poseur [27] (TP)	74.2	80.9	84.8	86.4	88.6	3.15
ED-Pose [43] (OS)	71.6	80.1	84.4	86.9	88.5	5.40
Click-Pose (OS)	**73.0**	**83.2**	**90.3**	**94.1**	**96.4**	**1.97**

Table 4: **Comparison with DETR-based models** on COCO `val` set.

Method	Backbone	AP	AP_M	AP_L
Model-Only				
ViTPose (TP)	ViT-Huge	28.7	1.6	31.8
HRNet (TP)	HRNet-w48	22.2	1.6	24.5
HrHRNet (BU)	HRNet-w48	34.6	5.6	38.1
ED-Pose (OS)	ResNet-50	37.5	7.6	41.1
Click-Pose-C0 (OS)	ResNet-50	40.5↑3.0	8.3	44.2
Model+Manual Correction				
ViTPose-C3	ViT-Huge	32.1	5.1	34.8
ViTPose-C5	ViT-Huge	36.1	12.3	38.3
ViTPose-C7	ViT-Huge	40.3	19.0	42.3
ViTPose-C9	ViT-Huge	47.5	28.9	49.1
Neural Interactive				
Click-Pose-C3	ResNet-50	61.6 (+13.4)	30.8 (+16.7)	65.1 (+13.3)
Click-Pose-C5	ResNet-50	71.8 (+19.8)	45.1 (+26.4)	74.5 (+19.2)
Click-Pose-C7	ResNet-50	78.5 (+24.1)	54.7 (+32.9)	80.9 (+23.3)
Click-Pose-C9	ResNet-50	**83.7** (+27.6)	**63.1** (+38.1)	**85.9** (+27.0)

Table 5: **Comparison with representative SOTAs** on Human-Art `val` set. All the models are trained on COCO and tested on Human-Art as out-of-distribution data.

Method	C0↑	C1↑	C2↑	C3↑	C4↑	NoC@95↓
Poseur [27] (TP)	21.2	23.1	24.9	26.4	28.0	12.19
ED-Pose [43] (OS)	37.5	40.1	42.0	43.3	44.3	9.88
Click-Pose (OS)	**40.6**	**47.1**	**54.9**	**61.6**	**67.1**	**6.45**

Table 6: **Comparison with DETR-based models** on Human-Art `val` set.

Pose by 2.5 AP on OChuman and 0.7 AP on CrowdPose in the *model-only* setting, showing its robustness in crowded scenes. Compared with *ED-Pose+manual correction*, *Click-Pose* exhibits significant improvements when receiving user clicks. This is because it not only adjusts the classification scores for all candidate predictions but also enhances localization accuracy. Both of these enhancements greatly contribute to improving annotation efficiency.

Method	AP	AP_{50}	AP_{75}	NoC@95
Model-Only				
ED-Pose (OS)	31.4	39.5	35.1	-
Click-Pose-C0 (OS)	33.9↑2.5	43.4	37.5	-
Model+Manual Correction				
ED-Pose-C1	33.0	39.6	35.5	13.50
ED-Pose-C2	33.7	39.6	35.6	
Neural Interactive				
Click-Pose-C1	83.0 (+46.4)	92.4 (+49.0)	88.0 (+49.0)	**1.93**
Click-Pose-C2	**90.9** (+52.4)	**96.3** (+52.5)	**93.3** (+53.4)	

Table 7: **Comparison with baseline models** on the crowded scene, where all the models are trained on COCO and tested on OCHuman `test` set.

Method	C0↑	C1↑	C2↑	C3↑	C4↑	NoC@95↓
ED-Pose (OS)	69.9	77.6	82.3	84.8	86.1	6.37
Click-Pose (OS)	**70.6**	**79.1**	**86.1**	**91.3**	**94.5**	**1.47**

Table 8: **Comparison with baseline models** on the crowded scene, where all the models are trained and tested on CrowdPose with defined 14 keypoints.

4.6. Ablation Study

Two Key Components. We evaluate the effectiveness of the proposed two key components on the COCO `val` set, as shown in Table 9. **First**, *Click-Pose* incorporates pose error modeling to enhance the self-correction ability of the model. Our results demonstrate that this training strategy can lead to a 1.2 AP improvement and reduce the convergence time from 60 to 40 epochs compared to the baseline model [43]. **Second**, the human-feedback loop training can also provide an additional improvement of 0.9 AP by enhancing the model's robustness.

Pose Error	Loop	AP	AP_M	AP_L	Epoch
		70.9	65.2	79.2	60e
✓		72.1	66.5	80.3	45e
✓	✓	**73.0**	**68.1**	**80.5**	**40e**

Table 9: Impact on **key components** of *Click-Pose*-C0.

Strategies	C2	C4	C6	C8
Random	84.2	90.1	94.1	96.5
Low score	88.5	93.0	95.6	97.6
Worse	**90.3**	**96.4**	**98.1**	**98.7**

Table 10: Ablation study on three **click strategies**.

Strategies	C2	C4	C6	C8
Only Once	90.1	95.2	97.4	98.6
Progressive	**90.3**	**96.4**	**98.1**	**98.7**

Table 11: Ablation study on two **loop strategies**.

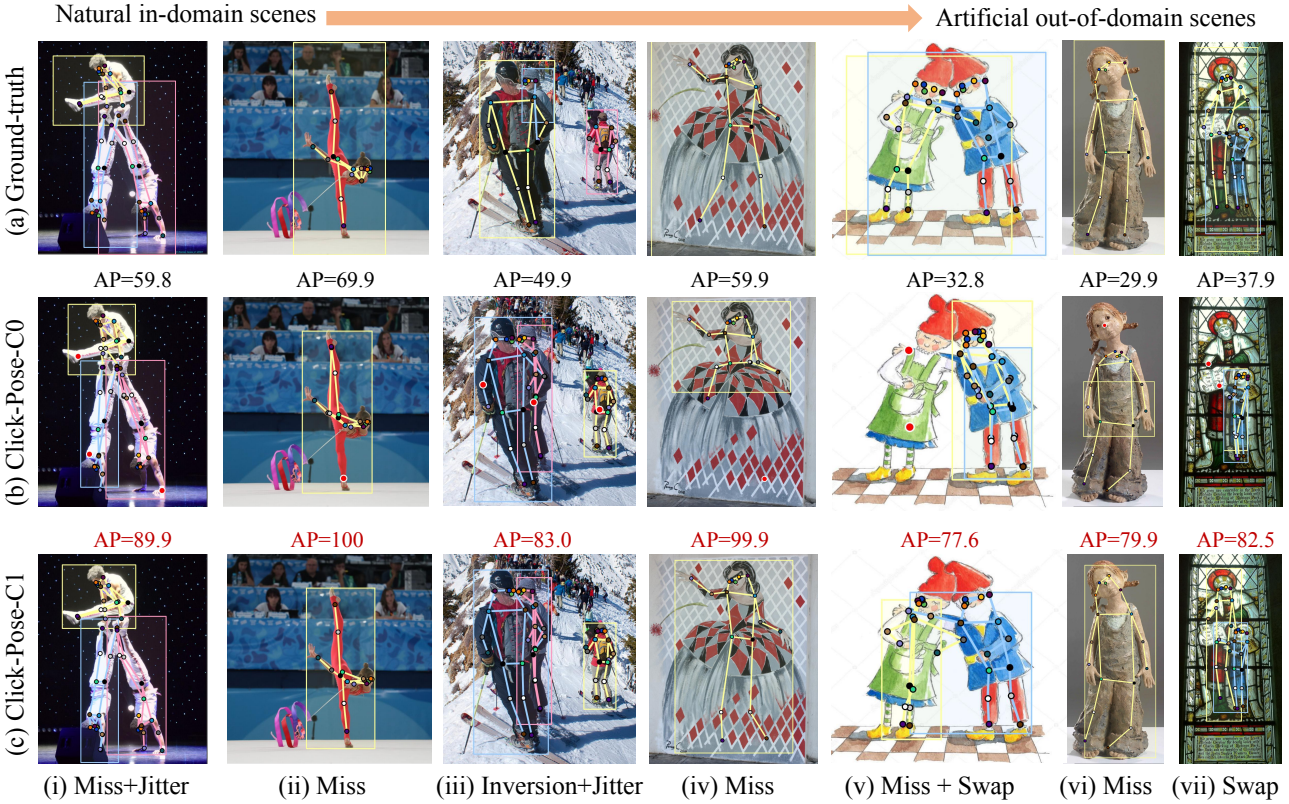


Figure 4: Visualization of the effects of the proposed *Click-Pose* with only **one** user click per person from in-domain scenes to out-of-domain scenes (directly test on them). The red dots in row (b) represent user clicks for *Click-Pose*-C0.

Method	AP@21	AP@17	AP@4	Correctable Range
Training on COCO				
Click-Pose-C0	25.1	40.5	0	-
*Interactive training with **100** annotated images*				
Click-Pose-C0	47.1$\uparrow_{22.0}$	52.0$\uparrow_{11.5}$	29.1$\uparrow_{29.1}$	-
Click-Pose-C2	58.2 (+5.4)	64.8 (+5.6)	33.2 (+4.1)	1-17
Click-Pose-C2	59.0 (+3.7)	61.1 (+3.3)	56.3 (+8.9)	1-21
*Interactive training with **1000** annotated images*				
Click-Pose-C0	55.0$\uparrow_{29.9}$	58.8$\uparrow_{18.3}$	40.9$\uparrow_{40.9}$	-
Click-Pose-C2	69.2 (+6.4)	**74.7** (+6.8)	45.4 (+4.5)	1-17
Click-Pose-C2	**70.4** (+5.9)	71.0 (+5.5)	**67.1** (+7.5)	1-21

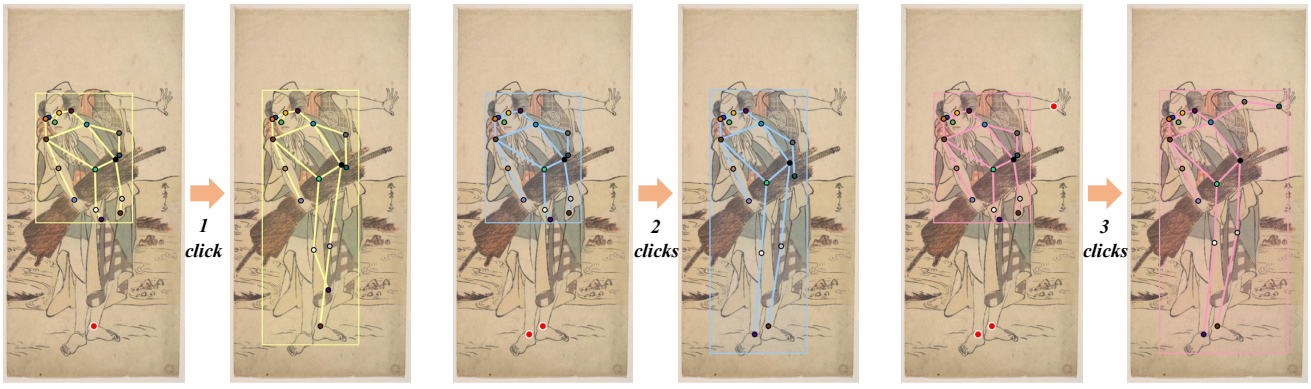
Table 12: Effect on **adaptation** 17 **to** 21 **keypoints** from COCO (17 keypoints) to Human-Art (21 keypoints).

Click Strategies. In experiments, we take the users to correct a worse keypoint by default. Besides, we explore two other strategies: random clicking and clicking on the keypoint with a low confidence score. In Tab. 10, *Click-Pose* shows consistent improvement under all three click strategies. Correcting the worse keypoint is the most intuitive annotation way and yields the best performance.

Loop Strategies. During inference, we set the progressive loop by default, where we only modify one worst keypoint in each loop iteration. We also explore the effectiveness of directly modifying multiple keypoints in a single loop iteration. Tab. 11 gives the performance trends of the two strategies for different numbers of clicks, showing that using the progressive loop strategy can obtain great results, and it is also more intuitive and user-friendly in practice.

4.7. Adaptation to Different Keypoints

We investigate the adaptability of *Click-Pose* in handling additional keypoints that are not included in the original training dataset. Specifically, we train the model using COCO with 17 labeled keypoints and finetune it on a small set of images (e.g., 100 and 1000) in Human-Art labeled with 21 keypoints, including 4 additional keypoints that are not defined in COCO. Tab. 12 reports AP@21 for all 21 keypoints, AP@17 for 17 keypoints defined by COCO,



(a) Click-Pose-C1 (AP=**70.0**) (b) Click-Pose-C2 (AP=**82.9**) (c) Click-Pose-C3 (AP=**95.0**)

Figure 5: Visualization of the effects of the proposed *Click-Pose* with varying numbers of clicks (e.g., one to three). The AP of *Click-Pose*-C0 is **40.0** AP, while *Click-Pose*-C3 could refine its results based on three clicks to achieve **95.0** AP.

and AP@4 for 4 additional keypoints that require adaptation. Moreover, we provide two options for the range of correctable keypoints, *i.e.,* 1-17 and 1-21. The results demonstrate that when the user corrects keypoints within the 17 keypoints defined by COCO, *Click-Pose* can refine an additional 4 keypoints. Furthermore, when expanding the range of correctable keypoints to include all 21 keypoints, the AP@4 score is further improved. Importantly, *Click-Pose* with limited annotated images can improve 22.0 AP to about 30.0 AP for AP@21 without manual correction.

4.8. Qualitative Results

Fig. 4 illustrates the effectiveness of the proposed *Click-Pose* in both natural in-domain and artificial out-of-domain scenes when receiving only *one* user click. By leveraging the human feedback loop, *Click-Pose* can refine other incorrect keypoints and boxes with user interaction. We show four typical pose error corrections, indicating the effectiveness and efficiency of our proposed method. Moreover, Fig. 5 shows how *Click-Pose* achieves increasingly better results with increasing clicks in challenging scenarios. As the number of clicks increases from 1 to 3, the AP score dramatically increases from 40 to 95.

5. Conclusion and Future Work

Conclusion. This work proposes a novel interactive keypoint detection task incorporating a human-in-the-loop strategy and presents *Click-Pose*, an end-to-end neural interactive keypoint detector. *Click-Pose* introduces two key components: a pose error modeling scheme and an interactive human-feedback loop. By effectively combining the model with user clicks, *Click-Pose* reduces labeling costs by over ten times compared to manual annotation. We hope this work will benefit the community by highlighting the importance of interaction between models and users.

Future Work. This work mainly focuses on multi-person 2D human pose estimation. There are some potential directions for future work. (I) **Interactive Whole-body Annotation**: Our work simply considers the mainstream 17 or 21 body keypoints. When dealing with more complex and dense keypoints (e.g., 133 keypoints [12, 44]), annotating small and blurry areas, like hands and faces, presents greater challenges. Importantly, these densely labeled body parts often exhibit locally structured spatial relationships that can be leveraged, making the task of labeling dense keypoints quite promising. (II) **Interactive Multi-task Annotation**: Our work has focused on annotating human body keypoints and their potential assistance in annotating body boxes (please see supplementary material). Similar to recent SAM [17], a more exciting direction is combining annotations from different tasks (like 2D/3D pose estimation, body parsing, and textual descriptions). A unified model could extract shared features and use different branches to obtain user inputs and estimate various annotations. Changing one annotation could affect others, offering a versatile and comprehensive annotation approach. (III) **Interactive 3D Annotation**: Annotating 3D needs high-cost devices and complex processing. Could we annotate 3D information (e.g., point cloud, mesh, keypoints) in the 2D space effectively [18, 31, 37]? This is an intriguing opportunity to expand this approach into the 3D domains.

Acknowledgment

The work is partially supported by the Young Scientists Fund of the National Natural Science Foundation of China under grant No.62106154, by the Natural Science Foundation of Guangdong Province, China (General Program) under grant No.2022A1515011524, and by Shenzhen Science and Technology Program JCYJ20220818103001002 and by Shenzhen Science and Technology Program ZDSYS20211021111415025.

References

[1] Hamed H Aghdam, Abel Gonzalez-Garcia, Joost van de Weijer, and Antonio M López. Active learning for deep detection neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3672–3680, 2019. 3

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3, 4, 6

[4] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 3

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 3, 7

[6] Mickael Cormier, Fabian Röpke, Thomas Golda, and Jürgen Beyerer. Interactive labeling for human pose estimation in surveillance videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1649–1658, 2021. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Qi Feng, Kun He, He Wen, Cem Keskin, and Yuting Ye. Rethinking the data annotation process for multi-view 3d pose estimation with active learning and self-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5695–5704, 2023. 3

[9] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021. 3

[10] Jia Gong, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Meta agent teaming active learning for pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11079–11089, 2022. 3

[11] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019. 3

[12] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2020. 1, 9

[13] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[14] Zhehan Kan, Shuoshuo Chen, Zeng Li, and Zhihai He. Self-constrained inference optimization on structural groups for human pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 729–745. Springer, 2022. 3

[15] Jinhee Kim, Taesung Kim, Taewoo Kim, Jaegul Choo, Dong-Wook Kim, Byungduk Ahn, In-Seok Song, and Yoon-Ji Kim. Morphology-aware interactive keypoint estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 675–685. Springer, 2022. 3

[16] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8166–8175, 2021. 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 9

[18] Theodora Kontogianni, Ekin Celikkan, Siyu Tang, and Konrad Schindler. Interactive object segmentation in 3d point clouds. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2891–2897. IEEE, 2023. 9

[19] Theodora Kontogianni, Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Continuous adaptation for interactive object segmentation by learning from corrections. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 579–596. Springer, 2020. 3

[20] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 5

[21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 5

[22] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1944–1953, 2021. 2

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 1, 5

[24] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4363–4372, 2017. 3

[25] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. *arXiv preprint arXiv:2210.11006*, 2022. 3, 5

[26] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13264–13273, 2021. 3

[27] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 72–88. Springer, 2022. 2, 7

[28] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7773–7781, 2019. 3, 5

[29] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 3

[30] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 369–378, 2017. 4, 5

[31] Tianchang Shen, Jun Gao, Amlan Kar, and Sanja Fidler. Interactive annotation of 3d object geometry using 2d scribbles. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 751–767. Springer, 2020. 9

[32] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022. 3, 4, 7

[33] Megh Shukla, Roshan Roy, Pankaj Singh, Shuaib Ahmed, and Alexandre Alahi. Vl4pose: Active learning through out-of-distribution detection for pose estimation. *arXiv preprint arXiv:2210.06028*, 2022. 3

[34] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 3, 5

[35] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[36] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 7

[37] Julien Valentin, Vibhav Vineet, Ming-Ming Cheng, David Kim, Jamie Shotton, Pushmeet Kohli, Matthias Nießner, Antonio Criminisi, Shahram Izadi, and Philip Torr. Semanticpaint: Interactive 3d labeling and learning at your fingertips. *ACM Transactions on Graphics (TOG)*, 34, 2015. 9

[38] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 2

[39] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8068–8078, 2022. 3

[40] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *European Conference on Computer Vision*, 2022. 2

[41] N. Xu, Brian L. Price, Scott D. Cohen, Jimei Yang, and Thomas S. Huang. Deep interactive object selection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 373–381, 2016. 3

[42] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 1, 2, 3, 6, 7

[43] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *International Conference on Learning Representations*, 2023. 2, 3, 4, 6, 7

[44] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. *arXiv preprint arXiv:2307.15880*, 2023. 9

[45] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation. In *European Conference on Computer Vision*. Springer, 2022. 3

[46] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *European Conference on Computer Vision*. Springer, 2022. 3

[47] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 5