

# CIGPose: Causal Intervention Graph Neural Network for Whole-Body Pose Estimation

## Supplementary Material

### A. Theoretical Foundation of Causal Intervention

This section provides the detailed theoretical support for the causal intervention framework introduced in the main paper Sec. 3. We first provide a formal derivation for the backdoor adjustment formula Eq. (1) using Pearl’s *do*-calculus [36] and then offer a theoretical justification for our counterfactual replacement approximation Eq. (3).

#### Derivation of the Backdoor Adjustment Formula.

Our goal is to estimate the interventional distribution  $P(Y|do(F))$ , which represents the true causal effect of the keypoint embeddings  $F$  on the final prediction  $Y$ , isolated from the confounding influence of visual context  $C$ .

We recall the Structural Causal Model (SCM) from Fig. 2(a) of the main paper. This model defines the causal paths  $C \rightarrow X \rightarrow F \rightarrow Y$  and the confounding paths  $C \rightarrow X$  and  $C \rightarrow Y$ . The critical issue is the non-causal backdoor path  $F \leftarrow X \leftarrow C \rightarrow Y$ , which allows spurious correlations between  $F$  and  $Y$  based on the confounder  $C$ .

To block this path, we must adjust for the confounding variable  $C$ . Using the three rules of *do*-calculus [36], we can formally derive the backdoor adjustment formula presented in Eq. (1).

Given a causal graph  $\mathcal{G}$ ,  $\mathcal{G}_{\overline{X}}$  denotes the graph with incoming edges to  $X$  removed, and  $\mathcal{G}_{\underline{X}}$  denotes the graph with outgoing edges from  $X$  removed.

- **Rule 1 (Insertion/deletion of observations):**

$$P(y|do(x), z, w) = P(y|do(x), w),$$

$$\text{if } (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}}}.$$

- **Rule 2 (Action/observation exchange):**

$$P(y|do(x), z, w) = P(y|x, z, w),$$

$$\text{if } (Y \perp\!\!\!\perp X|Z, W)_{\mathcal{G}_{\underline{X}}}.$$

- **Rule 3 (Insertion/deletion of actions):**

$$P(y|do(x), do(z), w) = P(y|do(x), w),$$

$$\text{if } (Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}, \overline{Z(W)}}}.$$

Our derivation proceeds as follows:

$$P(Y|do(F)) = \sum_c P(Y|do(F), c)P(c|do(F)) \quad (\text{A1})$$

$$= \sum_c P(Y|do(F), c)P(c) \quad (\text{A2})$$

$$= \sum_c P(Y|F, c)P(c) \quad (\text{A3})$$

**Step A1:** We begin by applying the law of total probability, marginalizing over the confounder  $C$ .

**Step A2:** We apply **Rule 3** (Insertion/deletion of actions) to show  $P(c|do(F)) = P(c)$ . This holds if  $(C \perp\!\!\!\perp F|\emptyset)_{\mathcal{G}_{\overline{F}}}$ . In the graph  $\mathcal{G}_{\overline{F}}$ , the edge  $X \rightarrow F$  is removed. The only paths from  $C$  to  $F$  are  $C \rightarrow X \rightarrow F$  (d-separated) and  $C \rightarrow Y \leftarrow F$ . The latter is a v-structure and is blocked since  $Y$  is not a condition. Thus,  $C$  and  $F$  are d-separated,  $(C \perp\!\!\!\perp F|\emptyset)_{\mathcal{G}_{\overline{F}}}$  holds, and the intervention  $do(F)$  has no effect on  $C$ .

**Step A3:** We apply **Rule 2** (Action/observation exchange) to show  $P(Y|do(F), c) = P(Y|F, c)$ . This holds if  $(Y \perp\!\!\!\perp F|C)_{\mathcal{G}_{\underline{F}}}$ . In the graph  $\mathcal{G}_{\underline{F}}$ , the edge  $F \rightarrow Y$  is removed. We must check if any other paths connect  $F$  and  $Y$ , given  $C$ . The only other path is the backdoor path  $F \leftarrow X \leftarrow C \rightarrow Y$ . Since we are conditioning on  $C$ , this path is blocked at  $C$ . Therefore,  $(Y \perp\!\!\!\perp F|C)_{\mathcal{G}_{\underline{F}}}$  holds, and we can replace the action  $do(F)$  with the observation  $F$ .

This completes the formal derivation of Eq. (1) in the main paper.

#### Justification for Counterfactual Replacement.

As stated in the main text, Eq. (A3) (the backdoor adjustment formula) is intractable to compute because  $C$  is unobserved, high-dimensional, and impossible to sum over. Our Causal Intervention Module (CIM) approximates this intervention by performing a *counterfactual replacement*,  $do(f_k := z_k)$ , as defined in Eq. (3).

The theoretical justification is as follows:

1. **Goal:** The target distribution  $P(Y|do(F)) = \mathbb{E}_c[P(Y|F, c)]$  represents the causal effect of  $F$  on  $Y$ , averaged over all possible contexts  $c$ . This conceptually creates a context-invariant, unbiased representation.
2. **Problem:** The observed embedding  $f_k \in F$  is confounded. It is a descendant of  $C$  via the path  $C \rightarrow X \rightarrow F$ . This dependency on  $C$  is precisely what opens the spurious backdoor path  $F \leftarrow X \leftarrow C \rightarrow Y$ .
3. **Our Approximation:** We introduce a learnable canonical embedding table  $Z \in \mathbb{R}^{K \times d_{emb}}$ . As a shared global parameter matrix,  $Z$  is reused across all images and is not conditioned on the current instance. Hence, it is optimized end-to-end while remaining independent of any specific input image  $X_i$  and its associated confounder  $C_i$ ; that is,  $Z \perp\!\!\!\perp C$  by construction.
4. **The Intervention:** When our CIM identifies a confounded embedding  $f_k$  (via  $s_c(k)$ ) and replaces it with its corresponding canonical ideal  $z_k \in Z$ , it is performing the counterfactual operation  $do(f_k := z_k)$ . This op-

eration replaces a variable that is dependent on  $C$  (the original  $f_k$ ) with a variable that is independent of  $C$  (the canonical  $z_k$ ).

5. **The Effect:** This replacement physically severs the causal link  $C \rightarrow X \rightarrow F$  at the feature level. For the intervened embedding  $f'_k = z_k$ , the backdoor path  $F' \leftarrow X \leftarrow C \rightarrow Y$  is broken because  $f'_k$  is no longer a descendant of  $C$  or  $X$ .

The shared/global parameterization of  $Z$  is central to this approximation. Each row  $z_k$  accumulates updates from many images whenever keypoint type  $k$  is selected for intervention, so sample-specific contextual cues are not tied to any single replacement vector. Instead, optimization reinforces what is consistently useful for predicting the same anatomical keypoint across diverse contexts, which is precisely the intended context-invariant prior. By computing  $P(Y|F')$  where  $F'$  is the deconfounded set, we force the model to reason using the context-invariant ideal  $z_k$  instead of the confounded evidence  $f_k$ . This serves as a practical, sample-specific approximation of the full, intractable summation over all contexts  $\sum_c$ . The counterfactual consistency loss defined in Eq. (6) further ensures that this replacement is targeted, regularizing  $Z$  to be a meaningful "ideal" representation for stable keypoints.

## B. Further Analysis of Predictive Uncertainty

As discussed in the main paper Sec. 1 and Sec. 3, our framework posits that predictive uncertainty is an effective proxy for identifying confounded keypoint representations. Confounders like heavy occlusion create a conflict between the model's spuriously learned priors from  $C \rightarrow Y$  and the visual evidence from  $F \rightarrow Y$ , resulting in high epistemic uncertainty.

**Beyond-occlusion quantitative evidence.** The occlusion-based analysis in Fig. 3 validates  $s_c(k)$  against a clearly defined confounder. To test whether the proxy also captures broader keypoint difficulty beyond explicit occlusion labels, we perform a within-instance top- $n$  enrichment analysis on the COCO-WholeBody validation set. For instance  $i$ , let  $V_i$  be the set of visible keypoints and let  $T_i = \text{TopK}(\{s_{c,i}(k)\}, n)$  be the keypoints selected by the proxy. We define

$$\Delta_i = \frac{1}{|T_i|} \sum_{k \in T_i} e_i(k) - \frac{1}{|V_i \setminus T_i|} \sum_{k \in V_i \setminus T_i} e_i(k), \quad (7)$$

$$e_i(k) = \|\hat{\mathbf{p}}_i(k) - \mathbf{p}_i(k)\|_2,$$

where  $e_i(k)$  is the localization error of keypoint  $k$ . A positive  $\Delta_i$  means that, within the same instance, the keypoints selected by  $s_c$  incur larger error than the remaining visible keypoints.

As shown in Tab. 5, the enrichment is strongly positive and widens on progressively harder subsets. This indicates

Table 5. Within instance top- $n$  enrichment on COCO-WholeBody. Keypoints selected by  $s_c$  exhibit larger localization error than the remaining keypoints in the same instance. Easy-drop  $p$  denotes the fraction of easiest instances removed when forming harder subsets, and 95% CIs are percentile bootstrap intervals over instances.

Easy-drop $p$	Kept inst.	Mean $\bar{\Delta}$ (px)	95% CI
0.00	104,125	8.47	[8.20, 8.72]
0.30	72,888	10.14	[9.78, 10.48]
0.50	52,063	10.45	[10.02, 10.87]

that  $s_c$  systematically concentrates on intrinsically harder keypoints within an instance, including cases driven by clutter, blur, and truncation rather than only binary occlusion status.

**Qualitative example.** In Fig. 3 of the main text, we quantitatively validated the proxy with occlusion labels. Here, in Fig. 7, we provide the qualitative example referenced in the main text.

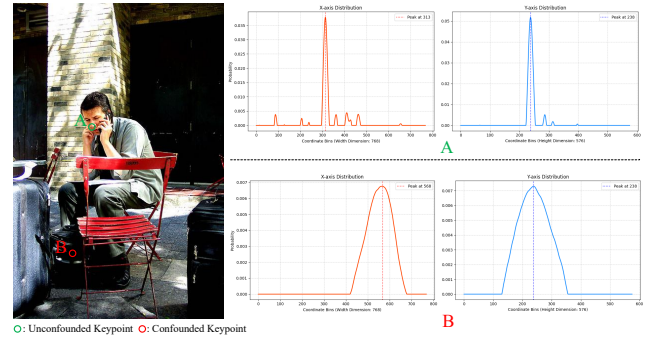


Figure 7. Visualization of posterior probability distributions. (A) For an unconfounded keypoint (green circle, nose), the distributions exhibit sharp, high peaks, indicating low uncertainty. (B) For a confounded keypoint (red circle, left ankle), which is occluded and in shadow, the distributions are diffuse with low peaks, signaling high predictive ambiguity.

This example visually demonstrates the mechanism of CIGPose:

1. **Confounder Identification:** The occluded L-Ankle leads to high predictive uncertainty, producing a diffuse posterior distribution.
2. **Causal Intervention:** Our model computes a high  $s_c(k)$  for this keypoint, triggering the counterfactual replacement  $do(f_{L-Ankle} := z_{L-Ankle})$ .
3. **Deconfounded Reasoning:** The Hierarchical GNN, now operating on the "clean" embedding set  $F'$ , leverages anatomical constraints from visible keypoints (such as the L-Knee) to infer the position of the occluded joint, resulting in an anatomically plausible prediction.

## C. Expanded Implementation Details

This section provides additional details on network architecture and training settings, supplementing Sec. 3 and Sec. 4.2 of the main paper.

**Network Architecture Details.** Our CIGPose framework builds upon the RTMPose [16] architecture, introducing a novel prediction head. Below, we detail the components of this head.

- **Backbone and Feature Processing:** We employ a CSP-NeXt backbone as our keypoint encoder. For an input image of size  $H \times W$ , the backbone outputs a feature map of size  $\frac{H}{32} \times \frac{W}{32}$  with 1280 channels. This feature map is then processed by a Gated Attention Unit (GAU) with hidden dimensions of 512, which extracts the initial keypoint embeddings  $F \in \mathbb{R}^{B \times K \times 512}$ , where  $B$  is the batch size and  $K$  is the number of keypoints ( $K = 133$  for the COCO-WholeBody dataset, and  $K = 17$  for the COCO and CrowdPose datasets).
- **Causal Intervention Module (CIM):** The CIM takes the initial embeddings  $F$  as input. First, two separate linear layers project  $F$  to generate initial 1D SimCC predictions  $(P_{k,x}, P_{k,y})$ . These are used to calculate the confounder score  $s_c(k)$  for each keypoint per Eq. (2). The module then identifies the top- $n$  most confounded embeddings and replaces them with corresponding vectors from a shared canonical embedding table  $Z \in \mathbb{R}^{K \times 512}$ . This table is implemented as a standard Embedding layer, initialized with a normal distribution  $\mathcal{N}(0, 0.01^2)$  and optimized end-to-end.
- **Hierarchical GNN:** The deconfounded embeddings  $F'$  are processed by a hierarchical GNN to model anatomical constraints, as discussed in Sec. 3.3.
  - **Intra-Part Modeling:** The first stage uses an EdgeConv [48] layer that operates over the standard anatomical skeleton graph  $\mathcal{G}_p$ . This layer models local kinematics by computing edge features (the difference between connected node features concatenated with the source node feature) and aggregating them via a shared MLP, which is implemented using a  $1 \times 1$  2D convolution.
  - **Inter-Part Attention:** The second stage uses an AttentionModule to capture long-range dependencies. It first computes an aggregate representation for each predefined semantic keypoint group by averaging the features of its constituent keypoints. These group-level features are then refined by passing them through a second EdgeConv layer defined over a fully-connected graph of all groups. The output is then used to generate channel-wise attention weights via a linear layer and a sigmoid function, per Eq. (4). These weights modulate the keypoint features, enabling the model to reason about global inter-part relationships.

The full skeleton graph definition and the list of semantic groups are specified in the model configuration files.

**Joint optimization of the canonical embedding table.** For a mini-batch of size  $B$ , the counterfactual replacement in Eq. (3) is applied instance-wise as

$$\begin{aligned} f'_{b,k} &= (1 - m_{b,k})f_{b,k} + m_{b,k}z_k, \\ m_{b,k} &= \mathbf{1}[k \in \mathcal{I}_b], \end{aligned} \quad (8)$$

where  $\mathcal{I}_b = \text{TopK}(\{s_c(b, k)\}_{k=1}^K, n)$  is the selected set for instance  $b$ , and the mask  $M = \{m_{b,k}\}$  is broadcast along the embedding dimension. Let  $H_\theta(\cdot)$  denote the prediction head following the encoder and hierarchical graph reasoning. The counterfactual path computes  $P_{\text{cf}}(Y) = H_\theta(F')$ , while the observational prediction  $P_{\text{obs}}(Y) = H_\theta(F)$  is used only through the stop-gradient target in Eq. (6). Therefore, gradients flow into the selected rows of  $Z$  according to

$$\frac{\partial \mathcal{L}}{\partial z_k} = \sum_{b=1}^B m_{b,k} \frac{\partial \mathcal{L}}{\partial f'_{b,k}}, \quad (9)$$

while rows that are not selected in the current mini-batch receive zero gradient from that step. Thus,  $Z$  is optimized jointly with the network parameters  $\theta$  under  $\mathcal{L} = \mathcal{L}_{\text{kpt}} + \lambda \mathcal{L}_{\text{cf}}$ .

The corresponding training procedure is summarized in Algorithm 1.

---

### Algorithm 1: Training: joint update of $(\theta, Z)$

---

**Input:** mini-batch images  $X$ ; GT distributions  $\{Q_{b,k}\}$ ; budget  $n$ ; weight  $\lambda$   
**Output:** updated parameters  $(\theta, Z)$   
 $F \leftarrow E_\theta(X)$  // keypoint embeddings  
 Compute  $(P_{b,k,x}, P_{b,k,y})$  and scores  $s_c(b, k)$  (Eq. (2))  
**for**  $b = 1$  **to**  $B$  **do**  
    $\mathcal{I}_b \leftarrow \text{TopK}(\{s_c(b, k)\}_{k=1}^K, n)$   
    $m_{b,k} \leftarrow \mathbf{1}[k \in \mathcal{I}_b], \forall k \in \{1, \dots, K\}$   
    $F' \leftarrow (1 - M) \odot F + M \odot Z$  // Eq. (3),  $M = \{m_{b,k}\}$   
    $P_{\text{cf}} \leftarrow H_\theta(F')$   
    $\tilde{P}_{\text{obs}} \leftarrow \text{sg}[H_\theta(F)]$  // stop-grad target  
    $\mathcal{L}_{\text{kpt}} \leftarrow \sum_{b,k} w_{b,k} D_{\text{KL}}(Q_{b,k} \parallel P_{\text{cf}}(Y_{b,k}))$   
    $\mathcal{L}_{\text{cf}} \leftarrow \frac{1}{\sum_b |\mathcal{S}_b|} \sum_b \sum_{k \in \mathcal{S}_b} D_{\text{KL}}(\tilde{P}_{\text{obs}}(Y_{b,k}) \parallel P_{\text{cf}}(Y_{b,k}))$   
   //  $\mathcal{S}_b$ : stable (non-intervened) set  
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{kpt}} + \lambda \mathcal{L}_{\text{cf}}$ ; update  $(\theta, Z)$  by AdamW

---

**Training settings.** Our training settings largely follow that of RTMPose [16], including a two-stage fine-tuning approach. The key hyperparameters for our main experiments (CIGPose-x on COCO-WholeBody) are summarized in Tab. 6.

Table 6. Key training hyperparameters for the CIGPose-x model.

Hyperparameter	Value
Optimizer	AdamW
Base Learning Rate	$2 \times 10^{-3}$
Weight Decay	0.05
LR Schedule	Cosine Annealing
Warm-up Iterations	1000
Max Gradient Norm	35
Training Epochs	420
Stage-2 Epochs	150
$\lambda$ for $\mathcal{L}_{cf}$	0.1 (Ablated in Sec. E)
Train Batch Size	32 (per GPU)
Intervention $n$	13 (Ablated in Sec. E)

## D. Data Augmentation

We adopt a two-stage data augmentation strategy to enhance model robustness, corresponding to the two-stage training detailed in Sec. C.

- **Stage 1 (Epoch 1-270):** This stage uses aggressive augmentation. It includes random horizontal flipping, random half-body transforms, random bounding box transformations (scaling from 0.5x to 1.5x, rotation of  $\pm 90^\circ$ ), top-down affine transforms, and YOLOX-style HSV color jittering. To simulate heavy occlusions, we also apply Albumentation transforms including `Blur` ( $p=0.1$ ), `MedianBlur` ( $p=0.1$ ), and `CoarseDropout` ( $p=1.0$ , `max_holes=1`, `max_height=128`, `max_width=128`).
- **Stage 2 (Epoch 271-420):** For the final fine-tuning stage, the augmentation intensity is reduced to allow the model to converge on cleaner data. Specifically, the random bounding box transform is modified to perform only scaling and rotation (shift factor is set to 0), and the probability of applying `CoarseDropout` is lowered to 0.5.

## E. Additional Experimental Analysis

We conduct further experiments on the COCO-WholeBody [19] validation set to analyze key components of our framework, supplementing the main ablations in Sec. 4.4. Unless otherwise specified, these studies use the CIGPose-l ( $384 \times 288$ ) model.

**Analysis of Intervention Frequency.** To verify that our Causal Intervention Module effectively targets keypoints susceptible to confounding, we analyze the intervention frequency for different body parts on the COCO-WholeBody validation set. As shown in Tab. 7, the intervention rates are highest for feet, hands, and legs. These are the body parts most frequently affected by common visual confounders such as occlusion, motion blur, and truncation. This empirical result provides evidence that our uncertainty-based

proxy successfully identifies and targets the most confounded keypoints.

Table 7. Intervention frequency per body part on COCO-WholeBody. The module intervenes most often on extremities, which are most prone to confounding.

Body Part	Intervention Freq. (%)
Face	0.67
Torso	0.16
Arms	0.24
Hands	0.90
Legs	0.89
Feet	1.36

**Sensitivity to Consistency Loss Weight  $\lambda$ .** The counterfactual consistency loss,  $\mathcal{L}_{cf}$  which is defined in Eq. (6), is critical for ensuring that interventions are meaningful and that the canonical embeddings  $Z$  are learned effectively. We analyze the model’s performance while varying the loss weight  $\lambda$ .

Table 8. Effect of the counterfactual consistency loss weight  $\lambda$ .

$\lambda$	Whole-Body AP (%)
0	65.8
0.01	66.1
<b>0.1</b>	<b>66.3</b>
0.5	65.9

As shown in Tab. 8, removing the consistency loss entirely ( $\lambda = 0$ ) leads to a 0.5 AP drop. This demonstrates that without this regularization, the model struggles to learn stable and meaningful canonical embeddings. A larger weight ( $\lambda = 0.5$ ) also slightly degrades performance, likely by overly constraining the GNN and preventing it from fully leveraging the deconfounded features. Our chosen value of  $\lambda = 0.1$  provides the best trade-off.

**Analysis of Intervention Strategy and Parameter  $n$ .** All main experiments use a fixed ‘top- $n$ ’ intervention strategy during training. For comparison, we also evaluate a threshold-based strategy. Formally, given a predefined threshold  $\tau$ , the threshold strategy defines the set of keypoints selected for intervention,  $\mathcal{K}_{\text{conf}}$ , as:

$$\mathcal{K}_{\text{conf}} = \{k \mid s_c(k) > \tau\}$$

where  $s_c(k)$  is the confounder score defined in Eq. (2). This means we intervene on any keypoint whose uncertainty-based score surpasses the threshold.

The results in Tab. 9 show that the ‘top- $n$ ’ strategy is more effective. We hypothesize that this is because it pro-



Table 9. Comparison of intervention strategies during training. The fixed-budget ‘top- $n$ ’ strategy provides a more stable and effective training signal, with  $n = 13$  yielding the best result on COCO-WholeBody.

Intervention Strategy (Training)	Whole-Body AP (%)
Threshold $\tau = 0.7$	65.9
Threshold $\tau = 0.8$	65.8
Top- $n$ ( $n = 11$ )	66.1
<b>Top-<math>n</math> (<math>n = 13</math>)</b>	<b>66.3</b>
Top- $n$ ( $n = 15$ )	66.0

vides a more stable training signal by ensuring a fixed number of keypoints  $n$  are intervened upon in each iteration. The ‘threshold’ approach can be noisy, as the number of interventions can vary dramatically between samples, potentially making it more difficult for the model to learn effective canonical embeddings. We also ablate the value of  $n$  (for training) and find  $n = 13$  (10% of  $K = 133$  keypoints) to be optimal.

## F. Limitations

### Limitations of Causal Intervention via Replacement.

The Causal Intervention Module (CIM) approximates the intractable *do*-operation by identifying confounded embeddings  $f_k$ , as signaled by high uncertainty  $s_c(k)$ , and replacing them with context-invariant canonical embeddings  $z_k$ . While effective for occlusion, this  $do(f_k := z_k)$  substitution carries a risk of over regularization. The core assumption is that high uncertainty correlates with visual confounding. However, this assumption can be violated in cases of valid but statistically rare poses, such as the intimate head-to-head interaction shown in Fig. 8 (top). In such scenarios, the visual evidence  $f_k$  is anatomically correct but lies on the tail of the pose distribution. The model exhibits high epistemic uncertainty simply due to the sample’s out-of-distribution nature. Consequently, CIM may incorrectly flag these valid embeddings as confounded and replace them with  $z_k$ , which represents a mean or canonical ideal. This can regularize the prediction too aggressively, discarding correct fine-grained geometry in favor of a generic high-likelihood pose.

### Limitations of Uncertainty as a Confounding Proxy.

The second limitation relates to the proxy  $s_c(k)$  itself. Our quantitative validation in Fig. 3 and Tab. 5 shows that  $s_c(k)$  is effective at identifying ambiguity-induced difficulty, including occlusion and harder non-occlusion cases. However, it fails when the model falls victim to confident semantic errors. As illustrated in Fig. 8 (bottom), the model hallucinates a human skeleton on a street lamp. This error stems from spurious correlations learned dur-

ing training—specifically, associating vertical linear structures with limbs. Critically, the model does not view this as an ambiguous state; it predicts the false positive with high confidence (low uncertainty), likely due to the strong shape similarity fulfilling the top-down detector’s prior. Because  $s_c(k)$  remains low, the CIM mechanism is bypassed entirely. This highlights a fundamental boundary of our framework: it is adept at rectifying uncertain representations but lacks an inherent rejection mechanism for confident misidentification caused by strong contextual confusion.

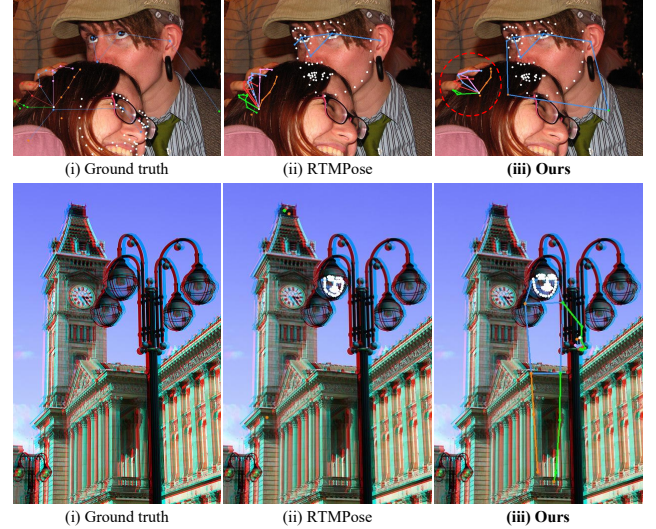


Figure 8. Visualization of failure cases.

## G. More Qualitative Results

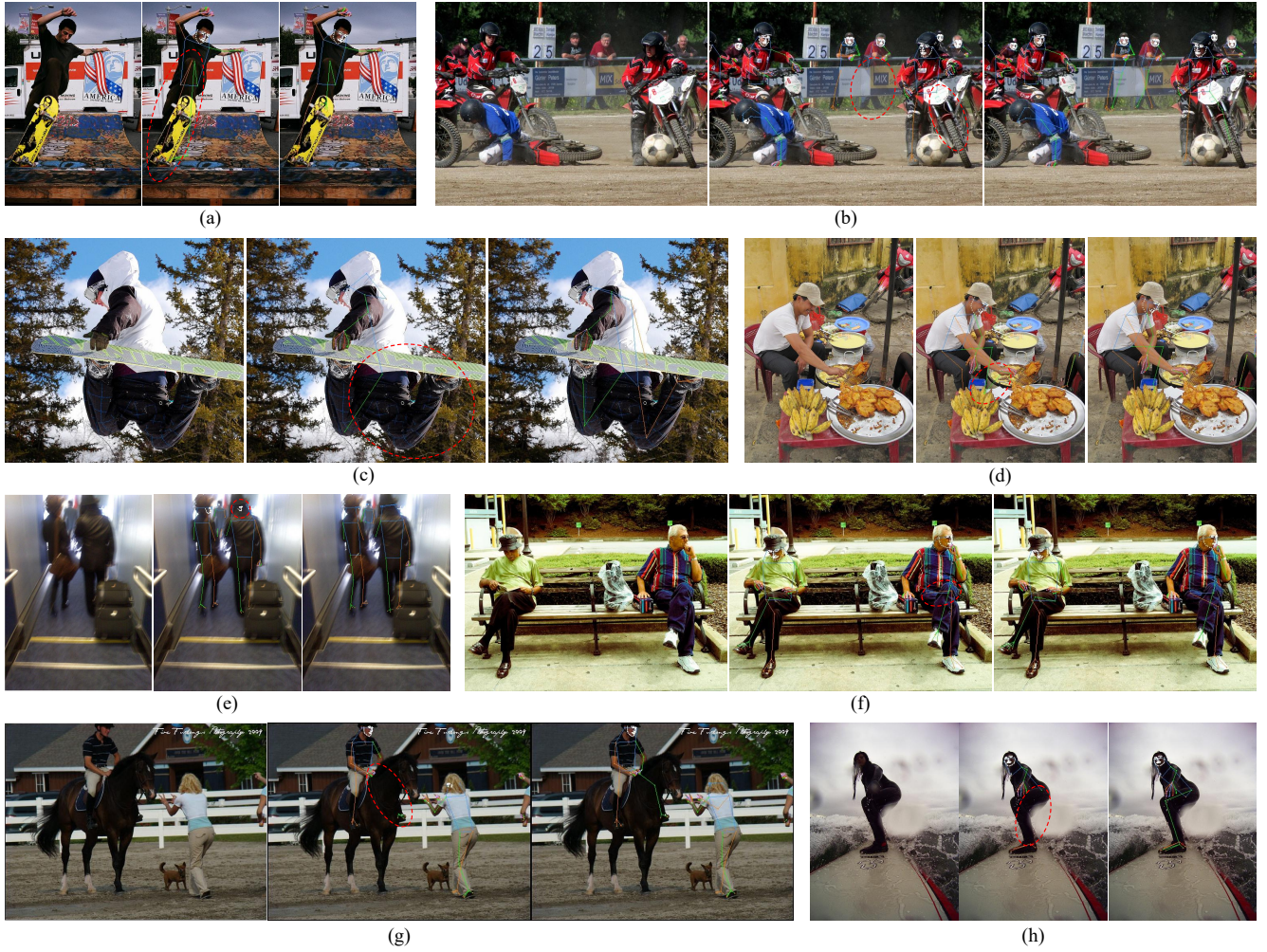


Figure 9. Qualitative comparison of CIGPose-x against the baseline RTMPose-x [16] on challenging images. From left to right: input image, RTMPose-x, and CIGPose-x.