Rethinking Visibility in Human Pose Estimation: Occluded Pose Reasoning via Transformers

Pengzhan Sun[1], Kerui Gu[1], Yunsong Wang[1], Linlin Yang[2], Angela Yao[1]
[1]National University of Singapore
[2]Communication University of China
{pengzhan, keruigu, yunsong, ayao}@comp.nus.edu.sg, lyang@cuc.edu.cn

Abstract

Occlusion is a common challenge in human pose estimation. Curiously, learning from occluded keypoints hinders a model to detect visible keypoints. We speculate that the impairment is likely due to a forced correlation between keypoints and visual features of the occluders. As such, we propose a novel visibility-aware attention mechanism to eliminate unreliable occluding features. The explicit occlusion handling encourages the model to reason about occluded keypoints using evidence and contextual information from the visible keypoints. It also mitigates the damage of unreliable correlations of the occluded keypoints. Our method, when added to the strong baseline SimCC, improves by 1.3 AP and 0.7 AP with ResNet and HRNet respectively. It also surpasses the state-of-the-art I^2R-Net on CrowdPose by 0.3 AP and 0.6 AP^{hard}. The improvements highlight that rethinking visibility information is critical for developing effective human pose estimation systems.

1. Introduction

Human pose estimation detects body keypoints in images and video. Deep learning methods [15, 18, 28, 34] are highly accurate at estimating pose, though occlusion remains an ongoing challenge. There are multiple sources of occlusion from objects, other people, and parts of the body itself. Under occlusions, there is no direct visual evidence for predicting keypoints, making the problem ambiguous.

The default in human pose estimation is to simply ignore the occlusions. Models are trained as if occluded keypoints were visible, based on speculated ground truth locations marked by annotators. Doing so forces models to correlate the visual evidence of the occlusion itself and the underlying keypoint. As we later show in Section 4.1, this type of correlation is spurious; in turn, these correlations harm models' ability to accurately predict *visible* keypoints.

A common approach to addressing occlusion is data augmentation [1, 23, 27]. The training data are augmented with simulated occlusions to encourages learned models to be more robust to the occlusions. However, this often requires specially designed training strategies and longer convergence times for training. For example, [11] doubles training epochs with only a 0.5 AP improvement on MSCOCO. Additionally, the gained robustness is rather limited. In fact, [24] demonstrates that existing data augmentation methods do not significantly improve the performance of top-down pose estimators and may even reduce performance for bottom-up approaches.

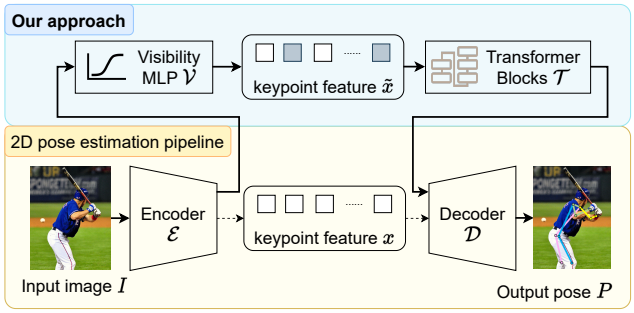

Figure 1. Our proposed module performs visibility-aware attention modeling on the keypoint representations to make them robust to occlusions. The module can be integrated into any pose estimation method with individual keypoint representations, such as heatmap-based methods [28,33] and CoordCls-based method [17].

A second line of work explicitly learns how to infer the invisible keypoints from image context. One common way is to use tailored architectures such as graph convolutional networks [25] and Siamese networks [37]. However, none of the existing works have explored the feasibility of directly using visibility labels as part of the training process. While such labels are not readily available for all datasets, they are provided in the commonly used MSCOCO [19].

To fill this gap, we first conduct observational studies that differentiate between visible and invisible keypoints for training. We compare the default mixed training with a vari-

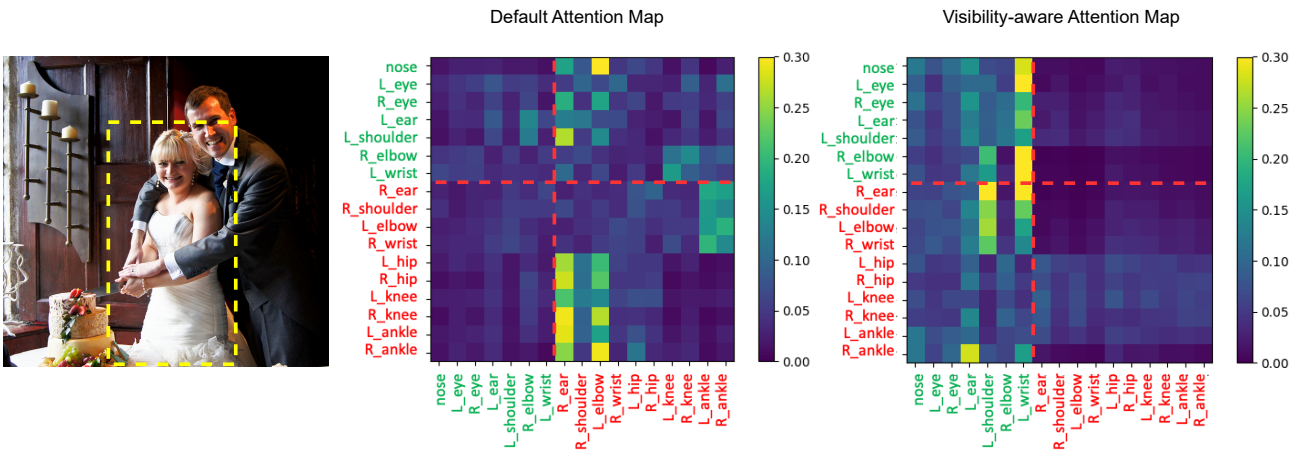

Figure 2. Human Keypoint Attention Map Comparison. x-axis represents keys and y-axis represents queries. Keypoints are grouped into visible keypoints and occluded keypoints. In the middle, the attention map is derived from a transformer with default mixed (visible and occluded) training, while the map on the right is produced by our visibility-aware transformer model. Note how our visibility-aware transformer suppresses the interference of occluded keypoints to other keypoints while enhancing the influence of visible keypoints.

ant where we train only with visible keypoints. Unsurprisingly, the default performs better on the occluded keypoints. Less expected was that the mixed default is less accurate than the visible-only variant, *i.e.*, training with only visible keypoints is more accurate. This phenomenon is especially pronounced on smaller datasets such as the MSCOCO `minitrain` [26] split (see Table 1 of Section 4.1).

Based on these observations, we propose a simple and effective module using avisibility-guided masking strategy. The module is plug-and-play (see Figure 1) and can be integrated into any pose estimation method which uses individual keypoint representations. Examples include the popular SimpleBaseline [33] and HRNet [28], which use explicit heatmap representations and SimCC [17], which use latent keypoint representations. After extracting keypoint features, we add a prediction network, called VisNet, to determine each keypoint's visibility. We then mask out occluded keypoint features and direct the resulting keypoint features, visible and occluded, into a transformer block before rerouting back to the main pose estimation pipeline. Masking features from the occluded keypoints, while simple, has the effect of forcing keypoints to be inferred from visual cues of only visible keypoints. It effectively guards against disturbances and interruptions from occluded keypoint features during attention calculations. For instance, Figure 2 shows how, in the attention map from the default method, the woman's visible left shoulder is affected by the occlusion on other keypoints, *e.g.*, her left elbow. To infer the woman's occluded right shoulder, the pose estimation model focuses mainly on unreliable occluded keypoints, *e.g.*, her left ankle and right ankle. After applying our explicit visibility masking, both visible and occluded keypoints pay more attention to visible keypoints, which benefits the overall keypoint accuracy.

Our proposed approach is highly effective on human pose datasets with both moderate [19] and extensive [16] occlusions. Summarizing our contributions:

1. We demonstrate that the default approach of including occluded keypoints on the same basis as visible keypoints for training negatively impacts model accuracy.
2. From this observation, we propose a novel plug-and-play module for occlusion handling. Our module uses visibility prediction and feature masking to guide models to rely only on cues from the visible keypoints.
3. Adding our module to state-of-the-art methods improves the accuracy of both visible and occluded keypoints. On SimCC [17] we improve on MSCOCO [19] by 1.3 AP; added to I^2R-Net [3] we improve by 0.6 AP^{hard} on the challenging CrowdPose dataset [16].

2. Related Work

2.1. 2D Human Pose Estimation

Human pose estimation is dominated by a top-down paradigm in which people are first localized in a scene with an external person detector before detecting the individual keypoints within the person bounding box. CNN- [17, 28, 30, 33] and transformer- [18, 20, 34, 35] based methods have steadily improved the overall pose accuracy. Different kinds of supervision signals and corresponding analysis for human pose estimation tasks have been explored. Previous works [10, 15, 17, 21, 29, 36] study different ways in which keypoint representations can be accurately decoded

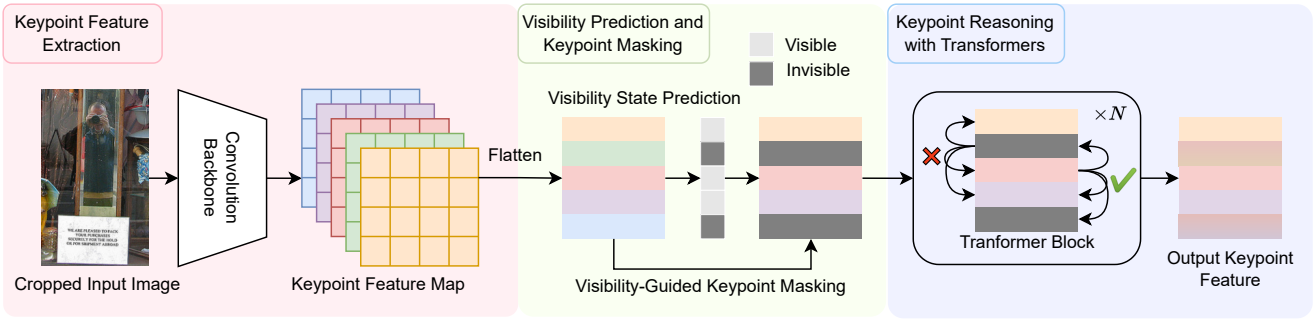

Figure 3. Model Architecture Overview. Our proposed framework comprises three integral components: Keypoint Feature Extraction, Masking, and Reasoning. The Visibility Prediction and Keypoint Masking module can apply current coordcls-based and heatmap-based pose estimation methods flexibly. The VisNet, employing an MLP, serves as a discriminative tool for binary visibility classification. The outcome of the Keypoint Reasoning with Transformers module can be seamlessly directed either to a linear coordinate classifier or, alternatively, transformed into heatmap features, thereby facilitating keypoint localization through an argmax function.

into keypoint locations. Others [5–7, 9, 13] have shown that predicting visible and occluded keypoints relies on different sources of information. Attribution maps [9] show that visible keypoints draw on localized visual cues while occluded keypoints source more global and dispersed cues. In this work, we take a step further to observe the impact of visibility and occlusion on human pose estimation. To our knowledge, we are the first to quantify the negative effect of occluded keypoint learning on visible keypoint accuracy. Specifically, we find that learning with occluded and visible keypoints indiscriminately reduces the accuracy of predicting visible keypoints.

2.2. Occlusion Handling

Methods for occlusion handling in human pose estimation can be divided into implicit and explicit methods. Implicit methods try to enhance robustness against occlusion using data augmentation. In [22], different frame crops were used to augment the number of partially observed samples. In [11, 12], different geometric patches overlaid on the keypoints in the image as augmentation for object occlusion. Augmentations against occlusion can also be applied at a feature level; for example, [23] uses an adversarial method to add occlusion patches on the feature map. The effectiveness of data augmentation methods depends on when they are introduced to training [11]. If not used properly, these methods suffer from an intrinsic drawback, which can have a detrimental effect on the model's performance [24].

Another line of work focuses on explicitly learning occluded keypoints. For instance, [2] uses attention modules to identify plausible clues from visible keypoints and estimate occluded keypoints. In [25], an image-guided graph convolutional network module was proposed to estimate invisible keypoints. [37] used knowledge distillation to transfer occlusion handling from one model branch to another of a Siamese network. Recent work [4] encodes human prior into the pose estimation system. However, these explicit occlusion handling methods only contribute to occluded keypoints' reasoning and do not actively mitigate the interference of learning from occluded keypoints themselves. In this work, we revisit the visibility information and propose utilizing the visibility flag as guidance to erase unreliable occluded keypoint features. This approach helps with both visible keypoint detection and occluded keypoint reasoning.

3. Preliminaries

3.1. A Primer on Human Pose Estimation

2D human pose estimation, also known as human keypoint detection, targets at detecting the locations of K human keypoints, *e.g.*, ankles, wrists, given an image input.

Keypoint Feature Encoding. State-of-the-art methods employ a top-down pipeline and extract individual keypoint features x from the given input image. Initially, a human-centric region I is cropped from the input image through a bounding box and subsequently resized to dimensions $(H, W, 3)$. An encoder $\mathcal{E}$, *e.g.*, a convolutional neural network (CNN), is subsequently deployed to extract keypoint features denoted as x, characterized by dimensions (K, H_f, W_f, C). Here, K represents the count of keypoints within each human pose, while H_f, W_f and C pertain to the height, width and channel number of the feature map corresponding to each keypoint.

Heatmap-based Decoding. A common and popular way to estimate human pose is through heatmaps. Heatmap-based methods transform this problem to estimating K heatmaps of size $W' \times H'$, $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K\}$, where each heatmap $\mathbf{H}_k$ indicates the location confidence of the kth keypoint. Given the extracted keypoint features x, heatmap-based methods firstly estimate K heatmaps $\{\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_K\}$ and then utilize the argmax function

Table 1. Observation study on the effect of including occluded keypoints in the training. We additionally distinguish the PCK on visible (PCK-Visible, 89%) and occluded keypoints (PCK-Occluded, 11%).

Training Split	Training Strategy	AP	AR	PCK-All	PCK-Occluded	PCK-Visible
`minitrain` [26]	all keypoints	57.5	61.7	77.7	**47.2**	81.9
`minitrain` [26]	visible keypoints	60.6	64.2	79.0	44.7	**83.4**
train2017	all keypoints	69.3	72.6	84.0	**55.2**	88.0
train2017	visible keypoints	68.8	72.1	83.9	52.6	**88.3**

to decode heatmaps into keypoint locations, *i.e.* the human pose. Formally, we have

$$\mathcal{L}_{pose} := \sum_{k=1}^{K} \left\| \hat{\mathbf{H}}_k - \mathbf{H}_k \right\|^2, \tag{1}$$

where $\hat{\mathbf{H}}_k$ is the predicted kth keypoint heatmap, and $\mathbf{H}_k$ is the kth keypoint ground truth heatmap, which is generated by applying 2D Gaussian with a standard deviation of one pixel centered on the ground truth location.

CoordCls-based Decoding. Another recently proposed method is coordinate classification [17], which applies individual classifiers to estimate the x-axis and y-axis coordinates. The continuous coordinate values are quantized into discrete integers as classification labels, i.e., the coordinate of x-axis $c_x \in [1, N_x]$, the coordinate of y-axis $c_y \in [1, N_y]$, where N_x and N_y are the number of bins for horizontal and vertical axes, respectively. Given the estimated coordinates classification softmax probability O_x and O_y, and the corresponding ground truth coordinates classes Y_x and Y_y, the loss function for coordinate classification is based on the KL-Divergence loss [14] $\mathcal{D}_{KL}$:

$$\mathcal{L}_{pose} := \mathcal{D}_{KL}(O_x, Y_x) + \mathcal{D}_{KL}(O_y, Y_y). \tag{2}$$

4. Method

Mainstream methods for human pose estimation [17, 18, 28, 35] can accurately estimate common visible human poses, but suffer from occlusion. We first conducted an observation study to quantitatively explore the impact of occlusion on human pose estimation (see Sec.4.1). Based on the findings, we propose a module to predict the visibility of each keypoint and mask the occluded nodes in subsequent attention blocks (see Figure. 3). Such an approach has two aims. First, it prevents the visual cues from occluded keypoints from interfering with the estimation of visible keypoints. Secondly, it encourages the occluded keypoints to be estimated with the help of visual cues of the visible keypoints. Our method can be integrated into any method leveraging individual keypoint features, like heatmap-based methods [28, 33, 35] and CoordCls-based methods [17].

4.1. Observational Study on Occlusion

Accurately predicting visible keypoints depends on local information surrounding the keypoints. However, the available visual cues may not provide enough evidence for accurate prediction of occluded keypoints. Therefore, predicting occluded keypoints requires global visual context, such as information from nearby visible keypoints [9]. Existing studies have not analyzed the roles of keypoints with different visibility statuses. To address this gap, we conducted an experiment examining the impacts of visible and occluded keypoints on learning.

We compare two models trained and evaluated on the MSCOCO [19] dataset. The first predicts visible keypoints and ignores losses resulting from occluded keypoints during the training process. As a comparison, the second model uses both visible and occluded keypoints for training. Similarly, we use a smaller MSCOCO [19] dataset split, `minitrain` [26], to train our two models. We use the SimpleBaseline [33] architecture with ResNet50 [8] for this observational study.

Table 1 shows the results for models trained on the train2017 split. The model that focuses on visible keypoints achieves 88.3% accuracy in predicting correct visible keypoints, surpassing the PCK-visible score of the model trained on all keypoints (including both visible and occluded keypoints) by 0.3%. Remarkably, when using only visible keypoints, models trained on the `minitrain` split exhibit not only a higher PCK-visible (83.4% compared to 81.9%), but also an improved PCK-All (79.0% compared to 77.7%) compared with the model incorporating all keypoints.

From our investigation, we derive the following interesting observations: 1) Incorporating occluded keypoints into the learning process detrimentally impacts the accuracy of predicting visible keypoints; 2) This adverse impact on accuracy becomes more pronounced when working with smaller datasets, subsequently degrading overall performance.

4.2. Visibility Prediction and Keypoint Masking

We utilize a Multilayer Perception (MLP) termed VisNet $\mathcal{V}$, to capture the visibility status of individual keypoints

while introducing a minimal number of parameters (less than 1MB) and low additional computational costs. Given the flattened keypoint features as input, VisNet $\mathcal{V}$ predicts a logit score $\hat{y}$ for keypoint visibility given the input x. By employing a sigmoid function, we transform this score into a probability, where values nearing 0 indicate an invisible keypoint, while values nearing 1 signify a visible keypoint. Following the obtained logit score of each keypoint, we apply a classification threshold of 0.5 to obtain the visibility mask, which we denote as $M_i(\hat{y}_i)$. The visibility mask is an 1D vector with the length of K, where i represents the index number of the ith keypoint. Specifically, the visibility mask has a value of either 0.01 or 1:

$$M_i(\hat{y}_i) = \begin{cases} 1, & \text{sigmoid}(\mathcal{V}(x)) < 0.5, \\ 0.01, & \text{sigmoid}(\mathcal{V}(x)) \geq 0.5 \end{cases}. \quad (3)$$

Subsequently, we mask out the corresponding features of invisible keypoints by applying the visibility mask to the keypoint features. We denote the resulting masked features as $\tilde{x}$, which we obtain by element-wise multiplication of the visibility mask M and keypoint features x:

$$\tilde{x} = M \odot x. \quad (4)$$

The rationale behind such visibility-based masking is that we leverage the predicted keypoint visibility to soft-mask the keypoint features, which provides additional clues in the subsequent Transformer module and empowers the model to refine the invisible keypoints with cross-keypoint attention.

4.3. Keypoint Reasoning with Transformers

After masking out the occluded keypoint feature, we employ a transformer network [31] to infer occluded keypoints and refine visible keypoints based on the output of the visibility prediction and keypoint masking module. Our keypoint reasoning algorithm takes the visibility-guided masked feature $\tilde{x}$ as input, and outputs enhanced keypoint features by visibility-aware attention.

Without bells and whistles, we utilize naive transformer blocks in our model framework. Each block contains a multi-head self-attention module and a multi-layer perception module. The self-attention head can be formulated as:

$$SA\left(\tilde{x}\right) = \text{softmax}\left(\frac{\tilde{x}W_Q\left(\tilde{x}W_K\right)^T}{\sqrt{d_h}}\right)\left(\tilde{x}W_V\right), \quad (5)$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are linear projection layers, d is the dimension of each keypoint token, and d_h is set to d. The multi-head self-attention module combines h self-attention heads, and can be formulated as:

$$MSA(\tilde{x}) = \left[SA_1(\tilde{x}); SA_2(\tilde{x}); \ldots; SA_h(\tilde{x})\right] W_P, \quad (6)$$

where d_h is set to d/h, $W_P \in \mathbb{R}^{(h\cdot d_h)\times d}$ is linear projection layer. We utilize the residual learning strategy in our transformer block implementation, and layer normalization is used before every module.

4.4. Loss Function

To supervise the binary classification of visible and invisible keypoints, we compute $\mathcal{L}_{vis}$ using the standard cross-entropy loss between the predicted $\hat{Y}$ and true labels Y:

$$\mathcal{L}_{vis} := \text{CE}(\hat{Y}, Y). \quad (7)$$

The overall loss for the multi-task learning framework is denoted as

$$\mathcal{L} = \mathcal{L}_{pose} + \lambda\mathcal{L}_{vis}, \quad (8)$$

where λ is the loss weight for the visibility loss function and modulates the proportion of these two tasks.

5. Experiment

5.1. Implementation Details

In our observational study, we adapt training strategies based on dataset scales. For the MSCOCO [19] full split, both models start from a common 90th-epoch checkpoint and train for 50 epochs, with one model using visible keypoints and the other using all keypoints. On the `minitrain` [26] split, both models share a 50th-epoch checkpoint and continue training for 50 epochs using visible keypoints and all keypoints. For human pose estimation on MSCOCO, we train ResNet50 [8] for 140 epochs with learning rate reductions at the 90th and 120th epochs, and HRNet-W32 [28] for 210 epochs with rate reductions at the 170th and 200th epochs. In CrowdPose [16] Multi-Person Pose Estimation, we train TransPose-H [35] for 240 epochs, with a multistep optimization and learning rate decline every 50 epochs. We detach the gradient from the VisNet branch so that it does not affect the training of the feature extractor, and set λ to 1.

5.2. MSCOCO Keypoint Detection

Dataset. The MSCOCO dataset is a large and challenging 2D human pose estimation dataset, containing over 200,000 images and 250,000 person instances labeled with 17 keypoints, which includes both location and visibility annotations. We trained our models on the MSCOCO train2017 set, which consists of 57,000 images and 150,000 person instances, and evaluated them on the val2017 set, which includes 5,000 images. The `minitrain` split used in our observational study shares the same object instance statistics as the full training set. It contains 15,000 images and 34,916 person instances, which is around 1/4 of the human instances in the train2017 set.

Table 2. Performance comparisons on the MSCOCO validation set.

Method	Scheme	Backbone	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR
SimCC [17]	CoordCls	ResNet50	72.3	91.4	79.2	69.5	76.5	75.5
Ours	CoordCls	ResNet50	**73.6**	**92.5**	**80.4**	**70.9**	**78.4**	**76.9**
SimCC [17]	CoordCls	HRNet-W32	77.2	**93.5**	83.5	74.3	82.0	80.1
Ours	CoordCls	HRNet-W32	**77.9**	**93.5**	**84.6**	**75.2**	**82.4**	**80.7**

Table 3. Performance comparisons of visible and occluded keypoints on the MSCOCO validation set.

Method	Keypoints Evaluation	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR
SimCC [17]	Visible Keypoints	77.0	93.4	84.3	74.9	80.7	80.3
Ours	Visible Keypoints	**78.7**	**93.7**	**86.7**	**76.6**	**81.9**	**81.8**
SimCC [17]	Occluded Keypoints	65.0	78.9	65.8	64.7	70.4	72.4
Ours	Occluded Keypoints	**65.5**	**79.2**	**67.2**	**65.9**	**70.5**	**73.0**
SimCC [17]	Overall Keypoints	72.3	91.4	79.2	69.5	76.5	75.5
Ours	Overall Keypoints	**73.6**	**92.5**	**80.4**	**70.9**	**78.4**	**76.9**

Evaluation metric. We followed the standard evaluation metrics of average precision and recall scores used in previous papers [17, 28, 33] based on the Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i \exp\left(-d_i^2/2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)},$$

where d_i is the Euclidean distance between the estimated keypoint and the ground truth keypoint, s is the scale of the cropped person instance, k_i is a constant that controls falloff, v_i is the visibility identity, and δ is an indicator function. The visibility identity v can take on three values: $v = 0$ indicates that the keypoint is unavailable, typically because it is located outside the image; $v = 1$ indicates that the keypoint is occluded, but its specific position can be inferred; $v = 2$ indicates that the keypoint is clearly visible. To discern the performance enhancements specifically achieved on visible and occluded keypoints, we adapt the indicator function as follows: For evaluating visible keypoints, we employ $\delta\left(v_i = 2\right)$, while for assessing occluded keypoints, we use $\delta\left(v_i = 1\right)$.

Results. We evaluated our method on the MSCOCO validation set and compared it with the baseline method SimCC [17]. Table 2 shows that our method achieves a considerable improvement over the strong SimCC model across both ResNet50 and HRNet-W32 backbones for 1.3 AP and 0.7 AP, respectively. Our method achieved a significant improvement over previous strong baselines on MSCOCO, a dataset with a low occlusion proportion (occluded keypoints versus visible keypoints is 11.3% versus 88.7%). Our method explicitly mitigates the detriments of occluded keypoints at the stage of attention modeling and therefore improves the overall performance by a large margin.

We extend our experiments to distinguish the performance gain for both visible and occluded keypoints within the MSCOCO dataset [19]. Table 3 showcases that our approach yields a noteworthy improvement of 1.7 AP for visible keypoints and 0.5 AP for occluded keypoints, facilitated by the precise prediction of visibility information, which attains an accuracy rate of 89%[1].

5.3. CrowdPose Keypoint Detection

Dataset. The CrowdPose dataset contains 20K images and 80K persons labeled with 14 keypoints. We trained the model on the 10K images in the train split and evaluated its performance on the 2K images in the validation split. This dataset is even more challenging than MSCOCO, as the proportion of occluded keypoints is much higher.

Evaluation metric. We follow the definition of *Crowd Index* and difficulty levels in CrowdPose [16]. In the ith person's bounding box, the *Crowd Index* is defined as

$$\text{Crowd Index } = \frac{1}{n}\sum_{i=1}^{n}\frac{N_i^b}{N_i^a},$$

where N_i^a represents the number of keypoints belonging to the ith person, while N_i^b represents keypoints number belonging to other persons. N_i^b/N_i^a is the crowd ratio of the

[1] In contrast to keypoint detection evaluation, which excludes unavailable keypoints (those with a visibility flag of $v = 0$), visibility classification categorizes unavailable human keypoints as occluded. Consequently, in this scenario, the data distribution on MSCOCO for training visibility prediction comprises 57.2% visible and 42.8% occluded keypoints.

Table 4. Comparison with previous SOTA methods on CrowdPose dataset.

Strategy	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
SBL [33]	60.8	81.4	65.7	71.4	61.2	51.2
HRNet [28]	71.3	91.1	77.5	80.5	71.4	62.5
JC-SPPE [16]	66.0	84.2	71.5	75.5	66.3	57.4
OPEC-Net [25]	70.6	86.8	75.6	-	-	-
CID [32]	72.3	90.8	77.9	78.7	73.0	64.8
I^2R-Net [3]	76.3	93.5	82.2	83.2	77.0	67.4
I^2R-Net [3] + Ours	**76.6**	**93.6**	**82.3**	**83.4**	**77.3**	**68.0**

Table 5. Ablation study of Keypoint Reasoning module and Keypoint Masking strategy utilized in our method.

Backbone	Keypoint Reasoning	Keypoint Masking	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
ResNet50			72.2	91.4	79.2	69.6	76.5	75.6
ResNet50	✓		72.3	91.4	79.2	69.5	76.5	75.5
ResNet50	✓	✓	**73.6**	**92.5**	**80.4**	**70.9**	**78.4**	**76.9**

ith human instance. n is the total number of people in the image. *Crowd Index* is divided into three levels: easy (0-0.1), medium (0.1-0.8), and hard (0.81-1). Besides standard evaluation metrics of AP and AR scores, we also report AP^{easy}, AP^{medium}, and AP^{hard}.

Results. We evaluate our method on a more crowded and challenging scene, as shown in Table 4. The current state-of-the-art method, I^2R-Net, builds connections for each body part of the human from both inter-human and intra-human levels. After integrating our method into I^2R-Net with TransPose-H backbone, we enhanced inter-human modeling by empowering visibility awareness, explicitly guiding occluded keypoint learning from visible contextual information, and avoiding the detrimental effect of unreliable occluder appearance information on visible keypoint features. This leads to a 0.3 AP improvement. Furthermore, our method can significantly improve the metric AP^{hard} by 0.6 with the help of our explicit visibility-awareness reasoning transformer block. This shows that under severe occlusion levels, visibility information can be of great help and points to a promising way to further explore how to solve the occlusion problem.

5.4. Ablation Study

Effectiveness of visibility-guided masking strategy. We conducted an ablation study to demonstrate the contribution of the visibility-guided masking strategy to the final result. Table 5 shows that solely leveraging the transformer block to model mutual-keypoint dependency brings only trivial improvement. The main performance gain comes from our proposed visibility-guided masking strategy. Compared to solely using the transformer block, our visibility-guided masking strategy brings a noticeable 1.3 AP improvement.

Influence of the visibility accuracy. We also investigated how the accuracy of our predicted visibility flag affects model performance, as shown in Table 6. As expected, our method's performance is lower compared to the one using the ground truth visibility flag for training and inference. However, we can see that there is only a small gap on the AP metric, showing that our method does not rely on ground truth visibility identity during inference. The accuracy of our predicted visibility identity is around 89%, which also supports this as a feasible and promising way of utilizing visibility information for handling human pose estimation.

5.5. Qualitative Results

In the study, we compare the qualitative results of SimCC method with and without our visibility-guided masking strategy. The findings are presented in Figure 4. Our strategy performs better in situations where the scene is crowded and there is significant occlusion. For instance, in the picture present in the second column, the knee and ankle nodes of the two players overlap. The baseline method fails to detect the occluded human body keypoint correctly, while our visibility-guided method can correct this error.

6. Limitation and Future Work

Failure cases. We show failure cases with severe occlusion and rare pose in Figure 5. As shown in the right picture, the pose of a person lying down is rarely seen in the dataset. Additionally, many keypoints, *e.g.*, knees, hips, are occluded by the ankle keypoints. In such scenarios, visible keypoints are often insufficient for our model to effectively

Table 6. Ablation study of different masking strategies.

Strategy	Visibility Accuracy	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Ground Truth	100%	**73.7**	**92.6**	**80.4**	**71.2**	77.9	76.6
Visibility Prediction	89%	73.6	92.5	**80.4**	70.9	**78.4**	**76.9**

Figure 4. Qualitative results of naive transformer (first row) and our visibility-aware transformer (second row) for keypoints reasoning. We use red squares to highlight our improved prediction qualities when faced with strong occlusion.

Figure 5. Failure cases with severe occlusion scenario and rare poses.

reason about the occluded ones.

Future work. Visibility information is a promising solution for handling occlusion problems. However, not all 2D pose estimation datasets provide a ground truth visibility flag for training. For future work, we will explore the semi-supervised learning setting of visibility information. We believe that the transferability of VisNet will help the community better utilize visibility information. Additionally, we intend to extend our method to the hand pose estimation task that also suffers occlusion problems.

7. Conclusion

In this paper, we quantitatively observe for the first time that occluded keypoint points interfere with the prediction of visible keypoint points. Inspired by this observation, we propose a visibility-aware framework that explicitly models the keypoint visibility information that was previously overlooked by mainstream methods. We empower the model to learn the visibility attributes of keypoints, and consequently devise a Transformer block to mine the cross-keypoint attention. Our method can be easily plugged into existing state-of-the-art human pose estimators and lead to clear performance improvements on both MSCOCO and CrowdPose datasets, verifying the effectiveness of our method for further enhancing the keypoint detection.

Acknowledgements This research / project is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (STEM RIE2025 MOE-T2EP20220-0015).

References

[1] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *ICCV*, pages 1212–1221, 2017. 1

[2] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *CVPR*, pages 1831–1840, 2017. 3

[3] Yiwei Ding, Wenjin Deng, Yinglin Zheng, Pengfei Liu, Meihong Wang, Xuan Cheng, Jianmin Bao, Dong Chen, and Ming Zeng. Iˆ 2r-net: intra-and inter-human relation network for multi-person pose estimation. *arXiv preprint arXiv:2206.10892*, 2022. 2, 7

[4] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–671, 2023. 3

[5] Kerui Gu, Linlin Yang, Michael Bi Mi, and Angela Yao. Bias-compensated integral regression for human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[6] Kerui Gu, Linlin Yang, and Angela Yao. Removing the bias of integral pose regression. In *CVPR*, pages 11067–11076, 2021. 3

[7] Kerui Gu, Linlin Yang, and Angela Yao. Dive deeper into integral pose regression. In *International Conference on Learning Representations*, 2022. 3

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 5

[9] Qiyuan He, Linlin Yang, Kerui Gu, Qiuxia Lin, and Angela Yao. Analyzing and diagnosing pose estimation with attributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4821–4830, 2023. 3, 4

[10] Junjie Huang, Zheng Zhu, Feng Guo, and Guan Huang. The devil is in the details: Delving into unbiased data processing for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5700–5709, 2020. 2

[11] Junjie Huang, Zheng Zhu, Guan Huang, and Dalong Du. Aid: Pushing the performance boundary of human pose estimation with information dropping augmentation. *arXiv preprint arXiv:2008.07139*, 2020. 1, 3

[12] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *ECCV*, pages 713–728, 2018. 3

[13] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *CVPR*, pages 11127–11137, 2021. 3

[14] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 4

[15] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *CVPR*, pages 11025–11034, 2021. 1, 2

[16] 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. 2, 5, 6, 7

[17] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *ECCV*, 2022. 1, 2, 4, 6

[18] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *CVPR*, pages 11313–11322, 2021. 1, 2, 4

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 4, 5, 6

[20] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, and Zhibin Wang. Tfpose: Direct human pose estimation with transformers. *arXiv preprint arXiv:2103.15320*, 2021. 2

[21] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2

[22] Soonchan Park and Jinah Park. Localizing human keypoints beyond the bounding box. In *CVPR*, pages 1602–1611, 2021. 3

[23] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *CVPR*, pages 2226–2234, 2018. 1, 3

[24] Rafal Pytel, Osman Semih Kayhan, and Jan C van Gemert. Tilting at windmills: Data augmentation for deep pose estimation does not help with occlusions. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10568–10575. IEEE, 2021. 1, 3

[25] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *ECCV*, 2020. 1, 3, 7

[26] Nermin Samet, Samet Hicsonmez, and Emre Akbas. Houghnet: Integrating near and long-range evidence for bottom-up object detection. In *ECCV*, 2020. 2, 4, 5

[27] István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. How robust is 3d human pose estimation to occlusion? *arXiv preprint arXiv:1808.09316*, 2018. 1

[28] 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. 1, 2, 4, 5, 6, 7

[29] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, pages 529–545, 2018. 2

[30] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[32] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11068, 2022. 7

[33] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018. 1, 2, 4, 6, 7

[34] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 1, 2

[35] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *CVPR*, pages 11802–11812, 2021. 2, 4, 5

[36] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020. 2

[37] Lu Zhou, Yingying Chen, Yunze Gao, Jinqiao Wang, and Hanqing Lu. Occlusion-aware siamese network for human pose estimation. In *ECCV*, 2020. 1, 3