DiffusionRegPose: Enhancing Multi-Person Pose Estimation using a Diffusion-Based End-to-End Regression Approach

Supplementary Material

1. Comparison on COCO Test-dev.

We provide the performance of our proposed approach compared to other state-of-the-arts methods on the COCO test-dev2017 dataset. As evidenced by the results presented in Table 1, our proposed DiffusionRegPose approach demonstrates a substantial superiority over all bottom-up and one-stage methods. Furthermore, it outperforms specific top-down methods, including Mask R-CNN and PRTR (ResNet). However, it is essential to acknowledge that our method exhibits a minor gap compared to the PRTR with the HRNet-w32 as the backbone.

2. Network Parameters of Proposed Model

Our model f_θ consists backbone, encoder E, human-detection decoder D_H, human-to-keypoint token expansion module F_{H2K} and diffusion decoder D. Among these components, both the encoder E and the diffusion decoder D differ from those utilized in ED-Pose [13], whereas the remaining components are implemented in the same way. The distinct components are outlined in Table 2, wherein the time embedding is incorporated into the encoder E. The time embedding involves the scaling and shifting of extracted features. The self-attention (SA) and cross-attention (CA) mechanisms employed in the diffusion decoder D align with the $SA(\cdot)$ and $CA(\cdot)$ modules discussed in Section 3.

3. Evaluation on Computational Complexity

Table 3 presents an evaluation of the computational complexity in terms of network parameters, floating-point operations (FLOPs), and processing speed in frames per second (FPS) by comparing end-to-end frameworks like ED-Pose [13] and GroupPose [6] with our proposed DiffusionRegPose. As can be seen, the incorporation of the modules specified in Table 2 leads to a marginal increase in the number of network parameters and inference time. However, the computational amount (FLOPS) of our model is reduced compared to other approaches. Since we are not focusing on optimizing the network structure in this work, there is still space to further improve our DiffusionRegPose in the future research, e.g., by designing a more lightweight network architecture.

4. Training Process

For a better overview about the training process, we present its details in Algorithm 1.

5. Visualized Results on COCO

We present supplementary qualitative results on the COCO Val2017 dataset in Figure 1. It is evident that our proposed

Table 1. Comparisons with state-of-the-art methods on COCO test-dev2017 dataset. "†" symbolizes the flip test. "TD", "BU", and "OS" denote the top-down, bottom-up, and one-stage methods, respectively. "HM", "BR" and "KR" indicate adopting heatmap-based losses, human box regression losses and keypoint regression losses, respectively. All AP values are displayed in %. The 1st, 2nd and 3rd place are color coded for metrics with more than three distinct values.

Method			Ref	Backbone	Loss	AP	AP_{50}	AP_{75}	AP_M	AP_L
Non-End-to-End	TD	Mask R-CNN [3]	CVPR 17	ResNet-50	HM	63.9	87.7	69.9	59.7	71.5
		Mask R-CNN [3]	CVPR 17	ResNet-101	HM	64.3	88.2	70.6	60.1	71.9
		PRTR† [5]	CVPR 21	ResNet-101	KR	68.8	89.9	76.9	64.7	75.8
		PRTR† [5]	CVPR 21	HRNet-w32	KR	71.7	90.6	79.6	67.6	78.4
	BU	HrHRNet† [1]	CVPR 20	HRNet-w32	HM	66.4	87.5	72.8	61.2	74.2
		DEKR† [2]	CVPR 21	HRNet-w32	HM	67.3	87.9	74.1	61.5	76.1
		SWAHR† [7]	CVPR 21	HRNet-w32	HM	67.9	88.9	74.5	62.4	75.5
		LOGO-CAP† [12]	CVPR 22	HRNet-w32	HM	68.2	88.7	74.9	62.8	76.0
	OS	DirectPose [11]	- 19	ResNet-50	KR	62.2	86.4	68.2	56.7	69.8
		CenterNet† [14]	- 19	Hourglass-104	KR+HM	63.0	86.8	69.6	58.9	70.4
		FCPose [8]	CVPR 21	ResNet-50	KR+HM	64.3	87.3	71.0	61.6	70.5
		InsPose [9]	ACM MM 21	ResNet-50	KR+HM	65.4	88.9	71.7	60.2	72.7
End-to-End	OS	PETR [10]	CVPR 22	ResNet-50	KR+HM	67.6	89.8	75.3	61.6	76.0
		ED-Pose [13]	ICLR 23	ResNet-50	BR+KR	69.8	90.2	77.2	64.3	77.4
		GroupPose [6]	ICCV 23	ResNet-50	KR	70.2	90.5	77.8	64.7	78.0
		DiffusionRegPose	-	ResNet-50	BR+KR	70.6	90.5	78.6	64.9	78.4

Table 2. Parameters of distinct modules from ED-Pose.

Name	Network structure	Channel (in, out)
Time embedding	Sinusoidal Position Embeddings + Linear + GELU + Linear	(1, 1024)
	SiLU + Linear	(1024, 256)
Self-attention (SA)	Linear_Q	(2, 256)
	Linear_K	(2, 256)
	Linear_V	(2, 256)
Cross-attention (CA)	Linear_Q	(2, 256)
	Linear_K	(2, 256)
	Linear_V	(2, 256)

Table 3. Compasrison of model parameters, FLOPs, and FPS with backbone of ResNet-50 on NVIDIA 3090 GPU.

Method	Parameters (M)	FLOPs (G)	FPS
ED-Pose	48.06	276.51	7.25
GroupPose	46.95	281.60	10.56
DiffusionRegPose	49.38	272.77	6.15

Algorithm 1 Training DiffusionRegPose model f_θ

Input: image x, keypoint labels y_0, GT of human boxes b_0, box class labels c_0, total number of diffusion steps T, model f_θ containing backbone, encoder E, human-detection decoder D_H, human-to-keypoint token expansion module F_{H2K} and diffusion decoder D.
repeat
  Sampling step index: $t \sim \text{Uniform}(\{1, \ldots, T\})$
  Extracting feature: $x^{fea} = backbone(x)$
  Tokenized representation: $F = E(x^{fea})$
  Human box token: $F_H = D_H(F)$
  Human-to-keypoint token expansion: $F_{H2K} = D_{H2K}(F, F_H)$
  Adding noise: $y_t = q\left(y_t \mid y_0, \zeta\right)$
  Tokens for self-attention: $Q_{y_t}, K_{y_t}, V_{y_t} = MLP_{X\in(Qs,Ks,Vs)}(y_t)$
  $Q_{CK} = SA(Q_{y_t}, K_{y_t}, V_{y_t})$
  Tokens for cross-attention: $Q_{K_{SA}} = MLP_{Qc}(Q_{CK})$, $K_{F_{H2K}}, V_{F_{H2K}} = MLP_{X\in(Kc,Vc)}(F_{H2K})$
  $cKpts, cBox = CA(Q_{K_{SA}}, K_{F_{H2K}}, V_{F_{H2K}})$
  Decoding: $y'_t, b_t, c_t = D\left(cKpts, cBox\right)$
  Keypoint loss: $L_k = \left\|y'_t - y_0\right\|_1$
  Box loss: $L_h = \left\|b_t - b_0\right\|_1$
  Classification loss: $L_c = FocalLoss(c_t, c_0)$
  Update entire model f_θ with gradient descent step: $\nabla_\theta\left(L_k + L_h + L_c\right)$
until Converged

DiffusionRegPose framework adeptly handles keypoint estimation in occluded scenarios, as shown in the yellow dashed circle. Please kindly zoom in for the best viewing.

6. Keypoint Completion in Various Scenes

According to the keypoint completion method introduced in Section 3, we present more human instances with invisible keypoints completed in various scenes in Figure 2 to demonstrate its efficacy. Please kindly zoom in for the best viewing.

7. Diffusion Steps Evaluation

Compared to multi-step denoising, our one-step denoising achieves sufficient accuracy, as evidenced in Table 4. We present visualized pose estimation results at various denoising stages in Figure 3. Specifically, the blue dots illustrate the prediction outcomes for the right elbow. Our proposed method showcases a greater diversity and an expanded spectrum of potential predictions for obscured keypoints of the human body, in contrast to the distribution of the clustered right elbow predicted by the ED-Pose method. The objective of DiffusionRegPose is not to predict a single optimal pose but rather to approximate a set of poses that can effectively represent the posterior distribution.

Table 4. Study on denoising steps on CrowdPose test set.

Method	step	AP	AP_{50}	AP_{75}	FPS
ED-Pose	-	69.9	88.6	75.8	7.25
DiffusionRegPose	1	72.68	91.15	79.25	6.15
DiffusionRegPose	2	72.73	91.15	79.27	4.17
DiffusionRegPose	3	72.72	91.12	79.24	3.23

8. Evaluation of Different Backbones

The evaluation of our method using the Swin-L backbone on CrowdPose is provided, significantly surpassing the ED-Pose (Swin-L). This is consistent with the adoption of ResNet-50 as backbone.

Table 5. Compared one-stage methods using different backbones on CrowdPose test set.

Method	AP	AP_{50}	AP_{75}	AP_E	AP_M	AP_H	AP_b
ED-Pose (ResNet-50)	69.9	88.6	75.8	77.7	70.6	60.9	60.2
ED-Pose (Swin-L)	73.1	90.5	79.8	**80.5**	73.8	63.8	-
DiffusionRegPose (ResNet-50)	72.7	91.1	79.3	79.3	73.3	64.9	63.1
DiffusionRegPose (Swin-L)	**73.9**	**91.8**	**80.4**	79.7	**74.7**	**66.4**	**63.6**

References

[1] Bowen Cheng, Bin Xiao, Jingdong Wang, Humphrey Shi, Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5385–5394, 2019. 1

[2] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. *IEEE/CVF Conference on Com-*

Figure 1. Qualitative comparison of DiffusionRegPose (based on ResNet-50) (in the third row) with ED-Pose (based on ResNet-50) (in the second row) on COCO val2017. The original images are displayed in the first row for reference.

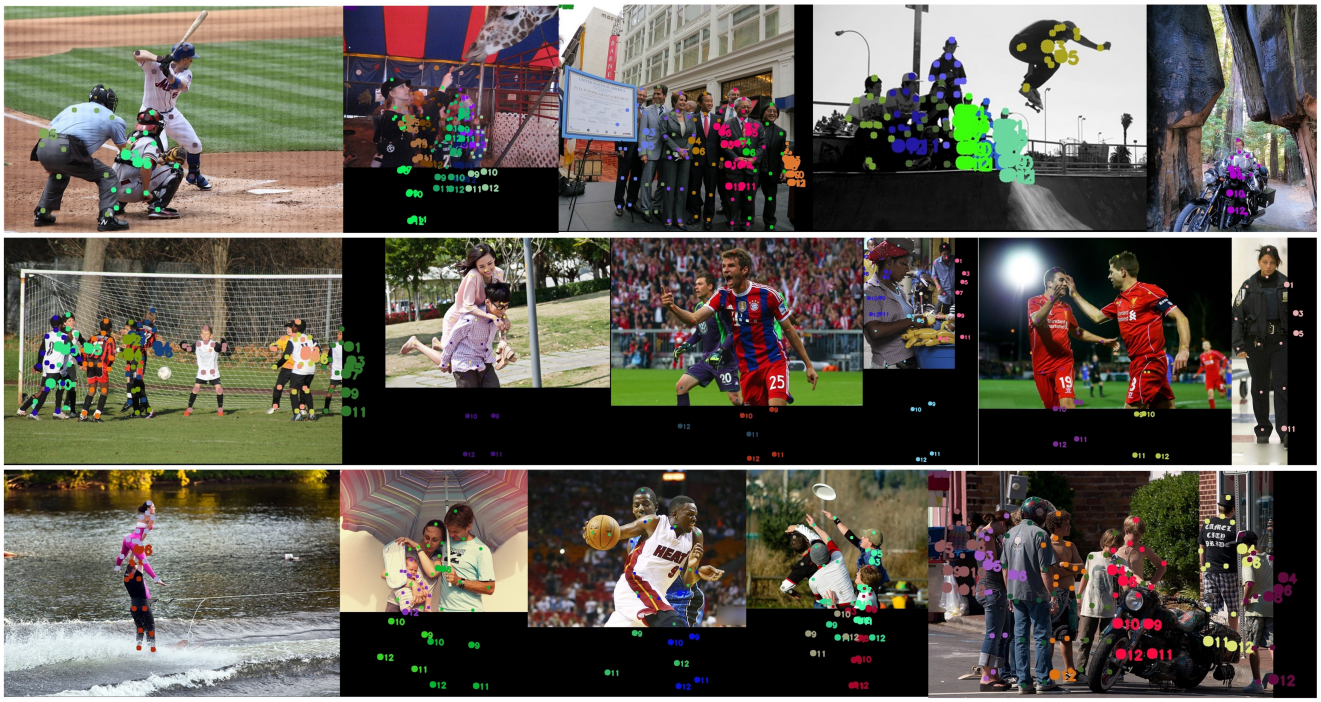

Figure 2. Invisible keypoint completion in various scenes. Notably, the completed invisible keypoint is displayed by a larger radius than the visible keypoint. The attached numerical value indicates the index of completed keypoint, such as index 11 for the left ankle. Additionally, diverse person instances are represented in distinct colors.

puter Vision and Pattern Recognition (CVPR), pages 14671–14681, 2021. 1

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397, 2017. 1

[4] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. *arXiv*, abs/2211.16487, 2022. 4

[5] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transform-

Figure 3. DiffusionRegPose produces poses with higher diversity, capturing the underlying uncertainty in a similar way to [4].

ers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1944–1953, 2021. 1

[6] Huan Liu, Qiang Chen, Zichang Tan, Jiangjiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, Yao Zhao, and Jingdong Wang. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 1

[7] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13259–13268, 2020. 1

[8] Wei Mao, Zhi Tian, Xinlong Wang, and Chunhua Shen. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9030–9039, 2021. 1

[9] Dahu Shi, Xing Wei, Xiaodong Yu, Wenming Tan, Ye Ren, and Shiliang Pu. Inspose: Instance-aware networks for single-stage multi-person pose estimation. *ACM International Conference on Multimedia*, 2021. 1

[10] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11059–11068, 2022. 1

[11] Zhi Tian, Hao Chen, and Chunhua Shen. Directpose: Direct end-to-end multi-person pose estimation. *arXiv*, abs/1911.07451, 2019. 1

[12] Nan Xue, Tianfu Wu, Gui-Song Xia, and L. Zhang. Learning local-global contextual adaptation for multi-person pose estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13055–13064, 2021. 1

[13] 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. 1

[14] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv*, abs/1904.07850, 2019. 1