PHASEPOSE: WAVELET PHASE DECOMPOSITION FOR POSE ESTIMATION ON LOW-RESOLUTION FEATURE MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate human pose estimation is commonly associated with restoring a high-resolution feature map. We show that this requirement is not intrinsic to the task. In a controlled degradation study, heatmap regression collapses when its output lattice is reduced from 64×48 to 8×6, whereas coordinate-based decoders remain near their full-resolution accuracy. The remaining obstacle is therefore not simply a small lattice, but how downsampling mixes sampling phases and how a decoder represents localization uncertainty. We propose PhasePose, an upsampling-free top-down estimator that addresses both issues. First, a critically sampled Haar analysis decomposes the four sampling phases of a stride-16 feature into one low-frequency and three directional-detail bands. The low-frequency band is kept fixed, while the three detail bands admit a grouped learnable $\mathrm{SO}(3)$ rotation before being stored as compact correction memory; no dense feature map is reconstructed. Second, a weight-shared spectral corrector represents each SimCC prediction by expectation, entropy, and a DCT-anchored spectrum whose analysis basis can adapt through a zero-start low-rank residual. Entropy can further condition a band-wise prior over the cached phase memory, while skeletal message passing and learned step embeddings support three recurrent gated logit corrections. Feature diagnostics show that the Haar anchor reduces mean inter-band correlation from 0.31 to 0.03, while the subsequently adapted directional-detail bands retain 48.1% of the illustrated response energy. On COCO val2017, PhasePose obtains 73.8 AP with ResNet-50 and 76.7/77.5 AP with CSPNeXt-L at 256×192/384×288. It also reaches 56.2 AP from a 64×48 input, 92.5 PCKh on MPII, 74.7 AP on CrowdPose, and 64.5 whole-body AP on COCO-WholeBody.

1 INTRODUCTION

Top-down human pose estimation localizes anatomical keypoints within detected person crops and remains a central 2D vision problem Chen et al. (2023). The prevailing solution predicts a dense heatmap for every joint. Stacked Hourglass, SimpleBaseline, RSN, and HRNet recover precise responses through repeated fusion, deconvolution, or sustained high-resolution representations Newell et al. (2016); Xiao et al. (2018); Cai et al. (2020); Sun et al. (2019). This design is effective because every heatmap cell has a direct spatial meaning, but it also couples coordinate precision to the output lattice. A coarse lattice cannot express nearby image coordinates as distinct peaks, so computation and activation memory are shifted to restored high-resolution stages Wang et al. (2023); An et al. (2024); Jiang et al. (2025). We revisit the underlying assumption: *must the feature map itself be restored before a precise pose can be decoded?*

The first question is what evidence should survive before the deepest feature becomes too coarse. A stride-32 feature is semantically strong, but stride-2 processing mixes four interleaved sampling phases in each 2×2 neighborhood; here, *phase* means these spatial offsets. Figure 1 shows that a fixed Haar anchor reduces their mean absolute off-diagonal correlation from 0.31 to 0.03. PhasePose keeps the low-frequency component fixed and rotates only the three detail components with grouped $\mathrm{SO}(3)$ transforms; the illustrated adapted details retain 48.1% of the response energy. Combined with uncertainty-aware correction, this yields a sharper right-ankle distribution without creating samples or increasing spatial resolution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

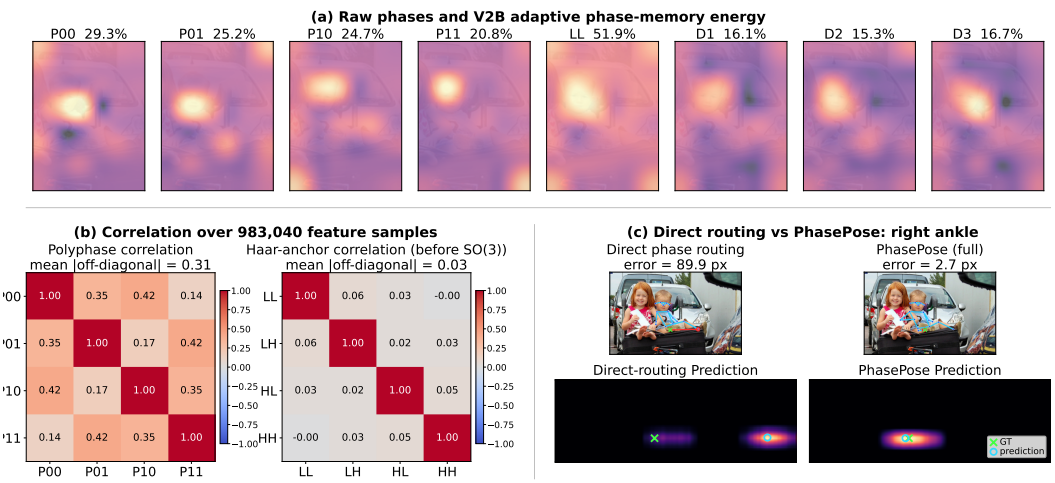

Figure 1: Why phase-resolved detail helps low-resolution pose correction. (a) Stride-16 sampling phases and Haar-anchored memory bands; the adapted detail bands retain 48.1% of the illustrated response energy. (b) Across 320 COCO crops, Haar reduces mean inter-band correlation from 0.31 to 0.03 before learned detail rotation. (c) A right-ankle example shows the sharper SimCC response after phase memory and recurrent spectral correction; green/cyan mark ground truth/prediction.

Having established the feature-side failure mode, we next ask whether a small lattice necessarily imposes imprecise coordinates. Starting from the same ResNet-50 He et al. (2016), we progressively remove deconvolution stages and expose Heatmap, SimCC Li et al. (2022), and SAR Wang & Zhang (2024) to 64×48, 32×24, 16×12, and 8×6 features. The methods are comparable at 64×48. Heatmap then falls from 71.8 to 17.2 AP, whereas SimCC and SAR remain between 70.8 and 72.5 AP. Their one-dimensional coordinate supports do not shrink with the feature lattice, so coarse spatial features do not impose coarse output bins. The complete degradation curve and model-scale bubbles are provided in the supplementary appendix.

Together, this controlled degradation and Figure 1 motivate the feature-side principle of PhasePose: use critically sampled wavelet analysis to preserve phase-resolved detail, but keep every subsequent spatial tensor on the stride-32 lattice. The Haar low-frequency band is fused with the deepest feature to form HW main tokens; the three detail bands are rotated only within their orthogonal subspace and retained as read-only correction memory. For a 256×192 crop, the main workspace contains only 48 spatial tokens. The method therefore differs fundamentally from wavelet super-resolution: it never synthesizes a denser image or feature map.

The remaining ambiguity lies in the output. Occlusion and weak distal evidence often produce broad, asymmetric, or multimodal coordinate distributions. An argmax suppresses these distinctions before refinement. PhasePose instead forms a compact approximation of each SimCC distribution from a DCT-anchored spectrum, expected coordinate, and entropy. The analysis basis is allowed to move only through a low-rank residual initialized at zero, and the entropy values can bias cross-attention toward the main or directional memory bands. A weight-shared corrector then propagates visually grounded evidence over the skeleton and adds a gated spectral residual to the logits; a learned embedding identifies each recurrent step without attenuating the correction magnitude. Thus, the two parts solve different failure modes: orthogonally routed phase detail protects local evidence from premature mixing, while recurrent distribution correction resolves uncertainty after coordinate decoding.

Our contributions are threefold:

- We separate feature resolution from localization resolution through a controlled degradation study, showing that the catastrophic low-resolution failure of heatmaps is primarily representation-dependent rather than an unavoidable property of pose estimation.
- We introduce Haar-anchored phase decomposition for low-resolution pose features. The low-frequency component remains analytic while grouped $\mathrm{SO}(3)$ rotations adapt the three-dimensional detail subspace, retaining correction evidence without spatial restoration.

- We formulate pose refinement as recurrent DCT-anchored correction from compact distribution-aware states, with low-rank spectral adaptation, entropy-conditioned memory reading, skeletal exchange, and shared recurrent weights. Together with extensive standard- and low-resolution experiments on four benchmarks, it produces a static, upsampling-free estimator with a favorable accuracy–efficiency trade-off.

2 Related Work

Heatmap estimators improve precision through repeated fusion, deconvolution, or sustained high-resolution features Newell et al. (2016); Xiao et al. (2018); Cai et al. (2020); Sun et al. (2019); DARK and UDP refine their discrete encoding and decoding Zhang et al. (2020); Huang et al. (2020), while recent methods make restoration lighter or sparser Wang et al. (2023); Jiang et al. (2025); An et al. (2024). Coordinate approaches instead decouple output bins from the feature lattice: residual likelihood regression, SimCC, attention-based keypoint tokens, RTMPose, and SAR progressively improve localization and efficiency Li et al. (2021a; 2022); Yang et al. (2021); Li et al. (2021b); Jiang et al. (2023); Lyu et al. (2022); Wang & Zhang (2024). ProbPose, visibility-aware reasoning, and CIGPose further model uncertainty or occlusion structure Purkrabek & Matas (2025); Sun et al. (2024); Li et al. (2026).

Plain transformers and continuous or pruned token representations provide complementary scaling strategies Xu et al. (2022b; 2023); Hu et al. (2024); Wang et al. (2024); lightweight, one-stage, and dynamically routed estimators target deployment efficiency Wang et al. (2022); Lu et al. (2024); Xu et al. (2025). Whole-body systems additionally use part-specific search, distillation, or heterogeneous data alignment Xu et al. (2022a); Yang et al. (2023); Jiang et al. (2024). Critically sampled wavelets provide a different primitive: complementary low- and high-frequency bands without increasing sample count Mallat (1989); Korkmaz et al. (2024). PhasePose uses Haar-anchored orthogonal phase routing rather than reconstruction and combines its adapted directional memory with recurrent DCT-anchored coordinate spectra on a fixed low-resolution graph.

3 Method

3.1 Overview

Figure 2 summarizes the pipeline. Let the backbone output

$$F_{16} \in \mathbb{R}^{B\times C_{16}\times 2H\times 2W}, \qquad F_{32} \in \mathbb{R}^{B\times C_{32}\times H\times W}. \tag{1}$$

PhasePose never increases H or W: it builds phase memory, predicts an initial SimCC distribution, and applies $T = 3$ shared corrections. At 256×192, $(H, W) = (8, 6)$ and $D = 256$, giving 48 main tokens and 192 memory tokens. At 64×48, the phase source shifts one stage earlier so PixelUnshuffle aligns a stride-8 map with the stride-16 main feature; stride-32 is pooled only as context.

3.2 Wavelet phase decomposition and memory

Polyphase decomposition. We first reduce F_{16} to $C_p = 64$ channels with a 1×1 convolution, batch normalization, and GELU. Let the reduced feature be $X = \phi_{16}(F_{16})$. Its four stride-2 polyphase components are

$$P_{ij}[:, :, h, w] = X[:, :, 2h + i, 2w + j], \qquad i, j \in \{0, 1\}. \tag{2}$$

PixelUnshuffle with factor two performs the same rearrangement without interpolation:

$$P = \mathcal{U}_2(X) = [P_{00}, P_{01}, P_{10}, P_{11}], \tag{3}$$

with $P_{ij} \in \mathbb{R}^{B\times C_p\times H\times W}$; it changes layout without inventing or discarding samples.

Haar-anchored orthogonal detail adaptation. For every channel and lattice location, the four phases are first multiplied by the fixed Haar matrix

$$\mathbf{H}_4 = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, \tag{4}$$

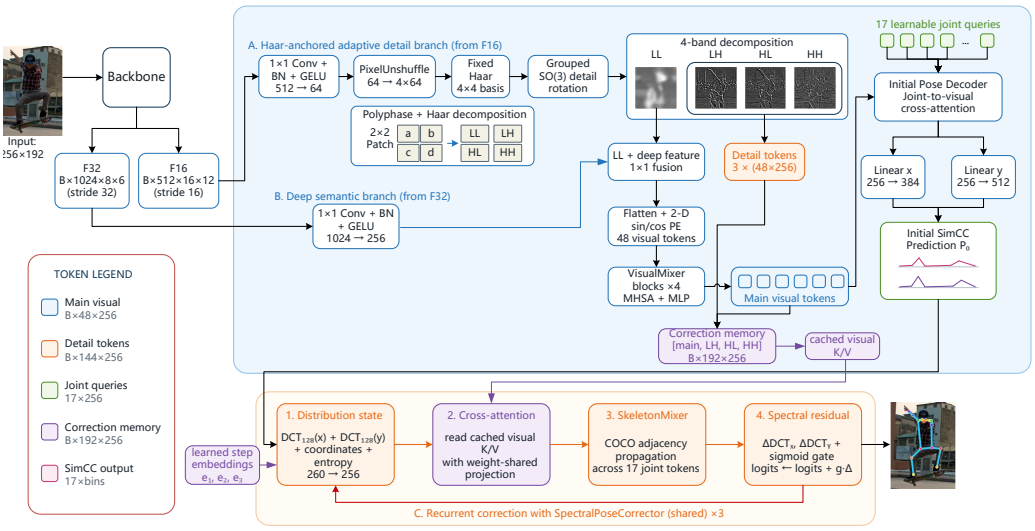

Figure 2: Overall architecture of PhasePose for a 256×192 input and CSPNeXt-L. The stride-16 feature is decomposed by a fixed Haar transform; the low-frequency band is fused with the stride-32 semantic feature, while grouped $\mathrm{SO}(3)$ rotations adapt only the three directional-detail bands. The resulting main and detail tokens form a compact correction memory. Joint queries first predict SimCC distributions, after which one weight-shared spectral corrector is applied for three recurrent steps using learned step embeddings.

$$[B_{LL}, B_1, B_2, B_3]^\top = \mathbf{H}_4 P. \tag{5}$$

Because $\mathbf{H}_4^\top \mathbf{H}_4 = \mathbf{I}$, Eq. (5) is an energy-preserving rotation of the four samples and is exactly invertible before projection. B_{LL} is the local average, while B_1, B_2, B_3 span the complementary directional-detail subspace. We keep B_{LL} fixed and allow only this three-dimensional detail subspace to adapt. Channels are divided into G groups; for group g, three learned scalars parameterize a skew-symmetric matrix $A_g \in \mathbb{R}^{3\times3}$ and an exact Cayley rotation

$$R_g = (I - A_g)^{-1}(I + A_g) \in \mathrm{SO}(3), \qquad [\widetilde{B}_1, \widetilde{B}_2, \widetilde{B}_3]_g = [B_1, B_2, B_3]_g R_g^\top. \tag{6}$$

The generators start at zero, hence $R_g = I$ initially. The low-frequency anchor stays fixed while the detail basis remains orthogonal within each group; grouped 1×1 projections map all bands to D channels.

The deep feature is projected to D channels and fused only with the low-frequency band:

$$M_0 = \mathrm{flat}(\phi_f[\phi_{32}(F_{32}); B_{LL}]) + E_{2D} + e_{LL}, \tag{7}$$

where E_{2D} is fixed two-dimensional sine/cosine encoding and e_{LL} is a learned band embedding. Four pre-normalized Transformer blocks produce the main visual tokens M. The correction memory is

$$\begin{aligned} V = [M; &\mathrm{flat}(\widetilde{B}_1) + E_{2D} + e_1; \\ &\mathrm{flat}(\widetilde{B}_2) + E_{2D} + e_2; \\ &\mathrm{flat}(\widetilde{B}_3) + E_{2D} + e_3]. \end{aligned} \tag{8}$$

Thus V stores four critically sampled $H \times W$ bands as $4HW$ tokens instead of reconstructing a $2H \times 2W$ map. The method changes how evidence is routed, not the number of samples.

3.3 Initial Pose Distribution

Let $Q \in \mathbb{R}^{K\times D}$ be K learnable joint queries. Cross-attention reads only the semantically mixed main tokens:

$$Q_0 = Q + \mathrm{Attn}(Q, M, M), \tag{9}$$

followed by an MLP and layer normalization. Two linear classifiers generate horizontal and vertical SimCC logits

$$Z_x^0 \in \mathbb{R}^{B\times K\times N_x}, \qquad Z_y^0 \in \mathbb{R}^{B\times K\times N_y}, \tag{10}$$

where $N_x = rW_{in}$, $N_y = rH_{in}$ and $r = 2$, so coordinate precision is decoupled from (H, W).

3.4 Shared spectral pose correction

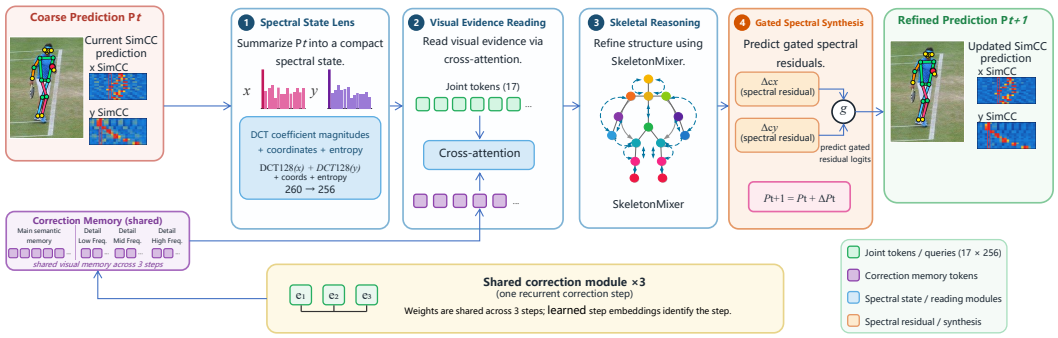

Figure 3: One recurrent spectral-correction step. The current SimCC prediction is summarized by DCT-anchored coefficients, coordinates, and entropy, projected to the joint-token state, and used to read the shared correction memory. SkeletonMixer then exchanges anatomically adjacent evidence before gated spectral residuals update the x/y logits. The module weights are shared across all three steps; learned embeddings e_1, e_2, e_3 identify recurrence depth.

Figure 3 summarizes one shared correction step. An argmax exposes only one bin and is unstable for ambiguous predictions. We instead retain the shape of each predicted distribution in a compact, DCT-anchored state. For axis $a \in \{x, y\}$ with N_a bins, let $p_a = \mathrm{softmax}(Z_a)$. The fixed orthonormal DCT-II analysis matrix $D_a \in \mathbb{R}^{N_a\times d_s}$ contains the first $d_s = \min(128, N_x, N_y)$ basis vectors,

$$[D_a]_{nk} = \eta_k \cos\left[\frac{\pi}{N_a}\left(n + \frac{1}{2}\right)k\right], \tag{11}$$

with $\eta_0 = N_a^{-1/2}$ and $\eta_k = (2/N_a)^{1/2}$ for $k > 0$. Rather than replacing this analytic basis, we parameterize a low-rank residual around it:

$$\widetilde{D}_a = D_a + \alpha_a U_a V_a, \qquad U_a \in \mathbb{R}^{N_a\times q},\ V_a \in \mathbb{R}^{q\times d_s}, \tag{12}$$

where q is small and the scalar α_a is initialized to zero. Thus the spectral analysis starts exactly from DCT and can learn only a low-rank deviation; setting $q = 0$ recovers the fixed analysis. The spectral coefficients are $c_a^t = p_a^t \widetilde{D}_a$. We further compute normalized expectation and entropy,

$$\mu_a^t = \sum_n p_{a,n}^t \xi_{a,n}, \qquad h_a^t = -\frac{1}{\log N_a}\sum_n p_{a,n}^t \log(p_{a,n}^t + \epsilon), \tag{13}$$

where $\xi_{a,n}$ is uniformly spaced in $[-1, 1]$. The state of each joint at step t is

$$s^t = [c_x^t, c_y^t, \mu_x^t, \mu_y^t, h_x^t, h_y^t] \in \mathbb{R}^{2d_s+4}. \tag{14}$$

Unlike an argmax, s^t preserves coarse distribution shape and uncertainty. The cap keeps it compact; at 64×48, $(N_x, N_y) = (96, 128)$ and $d_s = 96$.

The state is embedded and combined with a learned step embedding e_t before reading the shared memory:

$$\widetilde{Q}^t = Q^t + e_t + \phi_s(s^t). \tag{15}$$

The memory keys and values (K_V, V_V) are projected once from the immutable memory V and cached for all three steps. To let predictive uncertainty affect which phase bands are consulted, the reader optionally predicts four band probabilities from the current token and the two axis entropies,

$$\pi^t = \mathrm{softmax}\Big(\phi_b[\widetilde{Q}^t, h_x^t, h_y^t]\Big) \in \mathbb{R}^4. \tag{16}$$

For a memory token j belonging to band $b(j)$, the prior is injected directly into the attention logits,

$$A_{ij}^t = \mathrm{softmax}_j\left(\frac{q_i^t k_j^\top}{\sqrt{d}} + \log \pi_{i,b(j)}^t\right), \tag{17}$$

$$\bar{Q}^t = \widetilde{Q}^t + \mathrm{Proj}(A^t V_V), \tag{18}$$

$$Q^{t+1} = \bar{Q}^t + \phi_g(\widehat{A}\,\mathrm{LN}(\bar{Q}^t)). \tag{19}$$

The final layer of ϕ_b starts at zero, so π^t is initially uniform. $\widehat{A}$ is a row-normalized anatomical adjacency with self-loops; only this fixed graph changes across pose topologies.

Finally, the corrected tokens predict spectral coefficients $\Delta c_x^t, \Delta c_y^t \in \mathbb{R}^{d_s}$. Learnable synthesis bases $S_x \in \mathbb{R}^{d_s \times N_x}$ and $S_y \in \mathbb{R}^{d_s \times N_y}$ expand them to logit residuals. A per-joint sigmoid gate controls the update:

$$Z_a^{t+1} = Z_a^t + g^t(\Delta c_a^t S_a), \tag{20}$$

$$g^t = \sigma(w_g^\top \mathrm{LN}(Q^{t+1})), \quad a \in \{x, y\}. \tag{21}$$

Recurrence depth is represented by e_t without hand-designed attenuation. All corrector parameters are shared across $T = 3$ steps except the step embeddings, and inference remains a fixed-shape dense graph with no spatial interpolation or data-dependent iteration.

3.5 Training objective

Each prediction is supervised by the standard KL-based SimCC objective. Intermediate outputs encourage every shared step to remain predictive, while the final output also receives differentiable OKS supervision:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\mathrm{KL}}(Z^T, Y) \\ + \frac{\lambda_{int}}{T}\sum_{t=0}^{T-1}\mathcal{L}_{\mathrm{KL}}(Z^t, Y) \\ + \lambda_{oks}\mathcal{L}_{\mathrm{OKS}}(\mathbb{E}[Z^T], Y),\end{aligned} \tag{22}$$

where $\lambda_{oks} = 0.10$. The standard-resolution models use $\lambda_{int} = 0.10$, while the low-resolution transfer runs use $\lambda_{int} = 0.25$. Expectations for OKS are computed with softmax scale $\beta = 10$. The OKS area is transformed to the affine-crop coordinate system, ensuring that scale normalization is consistent with the model input.

4 Experiments

4.1 Protocol

We evaluate on COCO val2017 Lin et al. (2014), MPII Andriluka et al. (2014), CrowdPose Li et al. (2019), and COCO-WholeBody Jin et al. (2020). COCO uses train2017 and the common MMPose Faster R-CNN detections (56.4 human AP), with GT-box results separated. The COCO Keypoint Challenge test-dev2017 evaluation phase no longer accepts new submissions; because its annotations are withheld, new test-dev AP cannot be obtained, so we report COCO results on val2017. Unless noted, training lasts 210 epochs. All variants use $D = 256$, eight heads, four visual blocks, three shared correction steps, SimCC split ratio 2.0, DARK decoding, and $d_s = 128$ except $d_s = 96$ at 64×48. The CSPNeXt-L 256×192 setup uses four NVIDIA RTX 4090 GPUs with 256 crops per GPU (global batch 1024), AMP, and AdamW with learning rate 10^{-3} and weight decay 0.05. The backbone uses a $0.2\times$ learning-rate multiplier; normalization parameters and biases use zero weight decay; gradients are clipped to a maximum norm of 35. Unless noted otherwise, we select the best validation checkpoint and round to one decimal. Parameters/FLOPs use a single-crop graph; latency is batch-one FP32 eager inference. The full schedule, augmentation, initialization, and timing protocol are in the supplement.

Table 1: Top-down pose estimation on COCO val2017. The first three blocks use common MMPose Faster R-CNN detections (56.4 human AP); ‡ marks DynPose using its own Faster R-CNN boxes. GT-box results are separated. Red parentheses report our AP gain over NerPE, SHaRPose-S, ViTPose-B, and CIGPose-L in the corresponding blocks. CPU/GPU are batch-one FP32 forward-plus-decoding latency (ms); GPU uses eager execution. ⋆ denotes no public code.

Method	Source	Backbone	Input	Params (M)	FLOPs (G)	CPU (ms)	GPU (ms)	AP	AR
Detected boxes: ResNet-50 controlled comparison									
SimCC Li et al. (2022)	ECCV'22	ResNet-50	256×192	36.8	5.5	41.6	6.4	72.1	78.1
SAR Wang & Zhang (2024)	CVPR'24	ResNet-50	256×192	34.0	9.7	42.8	6.3	72.5	76.1
VA-SimCC⋆ Sun et al. (2024)	WACV'24	ResNet-50	256×192	–	–	–	–	73.6	76.9
NerPE Hu et al. (2024)	NeurIPS'24	ResNet-50	256×192	28.5	7.0	757.4	6.6	71.0	77.0
PhasePose (ours)	–	ResNet-50	256×192	29.7	4.3	34.4	6.6	**73.8** (+2.8)	**79.4**
Detected boxes: strong models at 256 × 192									
TokenPose-B Li et al. (2021b)	ICCV'21	TokenPose-B	256×192	13.5	5.7	99.8	15.9	74.7	80.0
HRNet Sun et al. (2019)	CVPR'19	HRNet-W32	256×192	28.5	7.7	87.2	20.8	74.9	80.3
ViTPose Xu et al. (2022b)	NeurIPS'22	ViT-S	256×192	24.3	5.6	36.8	9.8	73.9	79.2
MogaNet-S Li et al. (2024)	ICLR'24	MogaNet-S	256×192	29.0	6.0	71.2	12.8	74.9	80.1
DynPose‡ Xu et al. (2025)	CVPR'25	R50+HR32	256×192	63.4	6.8	69.3	14.3	74.8	–
RTMPose-L Jiang et al. (2023)	MMPose	CSPNeXt-L	256×192	27.7	4.2	40.6	10.7	75.8	80.6
SHaRPose An et al. (2024)	AAAI'24	ViT-S	256×192	28.4	4.9	47.9	11.1	74.2	79.5
PhasePose (ours)	–	CSPNeXt-L	256×192	32.0	4.4	46.3	10.2	**76.7** (+2.5)	**81.8**
Detected boxes: strong models at 384 × 288									
HRNet Sun et al. (2019)	CVPR'19	HRNet-W32	384×288	28.5	16.0	123.3	20.5	75.8	80.9
MogaNet-S Li et al. (2024)	ICLR'24	MogaNet-S	384×288	29.0	13.5	148.1	12.8	76.4	81.4
SHaRPose-B An et al. (2024)	AAAI'24	ViT-B	384×288	118.5	32.9	184.1	11.7	77.4	82.4
ViTPose-B Xu et al. (2022b)	NeurIPS'22	ViT-B	384×288	86.1	44.1	159.3	10.9	76.9	82.1
PhasePose (ours)	–	CSPNeXt-L	384×288	32.2	9.8	68.1	10.6	**77.5** (+0.6)	**82.3**
Ground-truth boxes									
ProbPose-S Purkrabek & Matas (2025)	CVPR'25	ViT-S	256×192	40.3	6.4	56.0	12.3	76.6	–
RTMPose-L Jiang et al. (2023)	MMPose	CSPNeXt-L	256×192	27.7	4.2	40.6	10.7	76.5	81.3
CIGPose-L Li et al. (2026)	CVPR'26	CSPNeXt-L	256×192	28.4	4.2	46.5	11.7	77.6	80.3
PhasePose (ours)	–	CSPNeXt-L	256×192	32.0	4.4	46.3	10.2	**78.5** (+0.9)	81.1
RTMPose-L Jiang et al. (2023)	MMPose	CSPNeXt-L	384×288	27.8	9.3	64.1	11.0	77.3	81.9
CIGPose-L Li et al. (2026)	CVPR'26	CSPNeXt-L	384×288	28.8	9.4	67.2	12.6	78.5	81.1
PhasePose (ours)	–	CSPNeXt-L	384×288	32.2	9.8	68.1	10.6	**79.7** (+1.2)	**82.0**

4.2 COCO ACCURACY AND EFFICIENCY

Table 1 compares accuracy and efficiency under matched box protocols. With ResNet-50, PhasePose reaches 73.8 AP while using 29.7M parameters and 4.3 GFLOPs. The CSPNeXt-L variants obtain 76.7/77.5 AP at 256 × 192/384 × 288, and the separate GT-box evaluations reach 78.5/79.7 AP. RTMPose is included as an MMPose-based efficiency reference.

Matched comparisons. PhasePose-R50 is 2.8 AP above NerPE (and 1.7 above SimCC) with 29.7M/4.3G parameters/FLOPs. At 256 × 192, PhasePose-CSPNeXt-L is 0.9 AP above RTMPose-L and reaches 81.8 AR. At 384 × 288, it is 0.6 AP above ViTPose-B while using 32.2M versus 86.1M parameters and 9.8 versus 44.1 GFLOPs. Under GT boxes it exceeds CIGPose-L by 0.9/1.2 AP at the two resolutions, separating pose quality from detector quality.

4.3 LOW-RESOLUTION INPUTS AND TRANSFER

Table 2 reduces the person crop under shared geometry. PhasePose-R50 improves SimCC/SAR by 4.6/4.7 AP at 64 × 48 and 2.3/2.4 AP at 128 × 96; CSPNeXt-L reaches 56.2/70.5 AP. The 64 × 48 models require only 0.4 GFLOPs. Coordinate classification avoids lattice quantization, while phase memory preserves evidence as the input contracts.

Transfer changes output topology; only joint queries, coordinate classifiers, and SkeletonMixer adjacency are dataset-specific. Table 3 keeps the same wavelet–spectral route across MPII, CrowdPose, and COCO-WholeBody.

Across topologies, PhasePose reaches 92.5 PCKh on MPII, 74.7 AP on CrowdPose, and 64.5 whole-body AP; per-part metrics are reported in the supplement.

Table 2: COCO val2017 at low pose-input resolutions with common detected boxes. Red parentheses report our AP gain over SimCC, CIGPose-L, SimCC, and CIGPose-L for the four PhasePose rows, respectively. CPU/GPU are batch-one FP32 forward-plus-decoding latency (ms); GPU uses eager execution.

Method	Backbone	Input	Params (M)	GFLOPs	CPU (ms)	GPU (ms)	AP
SAR Wang & Zhang (2024)	ResNet-50	64×48	34.0	0.5	16.6	6.1	39.2
SimCC Li et al. (2022)	ResNet-50	64×48	24.7	0.3	16.8	**5.8**	39.3
PhasePose (ours)	ResNet-50	64×48	29.6	0.4	**13.9**	6.2	**43.9** (+4.6)
HRNet Heatmap Sun et al. (2019)	HRNet-W32	64×48	28.5	0.5	32.4	20.5	29.7
UDP Huang et al. (2020)	HRNet-W32	64×48	28.5	0.5	32.9	21.1	47.4
DARK Zhang et al. (2020)	HRNet-W32	64×48	28.5	0.5	32.9	21.4	47.3
RTMPose-L Jiang et al. (2023)	CSPNeXt-L	64×48	27.5	0.3	21.1	6.0	53.0
CIGPose-L Li et al. (2026)	CSPNeXt-L	64×48	27.8	0.3	20.5	6.1	52.4
PhasePose (ours)	CSPNeXt-L	64×48	31.8	0.4	20.3	9.8	**56.2** (+3.8)
SAR Wang & Zhang (2024)	ResNet-50	128×96	34.0	1.8	25.6	**5.7**	62.5
SimCC Li et al. (2022)	ResNet-50	128×96	25.0	1.3	27.2	5.9	62.6
PhasePose (ours)	ResNet-50	128×96	29.5	**1.1**	**19.6**	6.0	**64.9** (+2.3)
HRNet Heatmap Sun et al. (2019)	HRNet-W32	128×96	28.5	1.9	48.5	20.4	61.5
DARK Zhang et al. (2020)	HRNet-W32	128×96	28.5	1.9	49.3	21.6	69.8
UDP Huang et al. (2020)	HRNet-W32	128×96	28.5	1.9	49.0	21.0	69.5
RTMPose-L Jiang et al. (2023)	CSPNeXt-L	128×96	27.5	1.0	26.9	10.2	69.6
CIGPose-L Li et al. (2026)	CSPNeXt-L	128×96	28.0	1.1	27.8	**10.4**	68.6
PhasePose (ours)	CSPNeXt-L	128×96	31.8	**1.1**	27.3	10.1	**70.5** (+1.9)

Table 3: Transfer across pose topologies. Gray rows are matched reference models; red parentheses show the PhasePose gain. $*$ denotes additional pose data and/or distillation.

(a) MPII validation

Method	Backbone	Input	Params (M)	GFLOPs	PCKh@0.5
TokenPose-L/D6 Li et al. (2021b)	HRNet-W48	256×256	21.5	13.0	90.2
HRNet Sun et al. (2019)	HRNet-W32	256×256	28.5	10.2	90.0
SimCC Li et al. (2022)	ResNet-50	256×256	23.6	5.4	90.0
SAR Wang & Zhang (2024)	ResNet-50	256×256	34.0	7.3	87.5
RTMPose-M* Jiang et al. (2023)	CSPNeXt-M	256×256	13.6	2.6	90.7
PhasePose (ours)	CSPNeXt-L	256×192	32.0	4.4	**92.5** (+1.8)

(b) CrowdPose test

Method	Backbone	Input	AP	AP^E	AP^M	AP^H
HRNet Sun et al. (2019)	HRNet-W32	256×192	71.3	80.5	71.4	62.5
HRFormer-B Yuan et al. (2021)	HRFormer-B	256×192	72.4	80.0	73.5	62.4
NerPE Hu et al. (2024)	HRNet-W32	256×192	66.9	74.2	68.0	56.4
CIGPose-L Li et al. (2026)	CSPNeXt-L	256×192	73.7	82.8	75.1	61.2
PhasePose (ours)	CSPNeXt-L	256×192	**74.7** (+1.0)	**83.9** (+1.1)	**76.2** (+1.1)	61.7 (+0.5)

(c) COCO-WholeBody validation at 256×192

Method	Backbone	Input	Params (M)	FLOPs (G)	Whole AP	Body AP	Foot AP	Face AP	Hand AP
TCFormerV2-S Zeng et al. (2024)	TCFormerV2-S	256×192	–	7.3	57.5	69.5	69.4	65.2	53.5
RTMPose-L Jiang et al. (2023)	CSPNeXt-L	256×192	28.4	4.5	61.1	69.5	65.8	83.3	51.9
CIGPose-L Li et al. (2026)	CSPNeXt-L	256×192	28.4	4.6	62.6	71.2	69.0	83.3	54.0
RTMW-M* Jiang et al. (2024)	CSPNeXt-M	256×192	32.3	4.3	58.2	67.6	67.1	78.3	49.1
DWPose-L* Yang et al. (2023)	CSPNeXt-L	256×192	28.4	4.5	63.1	70.4	66.2	**84.3**	**56.6**
PhasePose (ours)	CSPNeXt-L	256×192	32.0	5.0	**64.5** (+1.4)	**74.7**	**73.4**	81.6	54.0

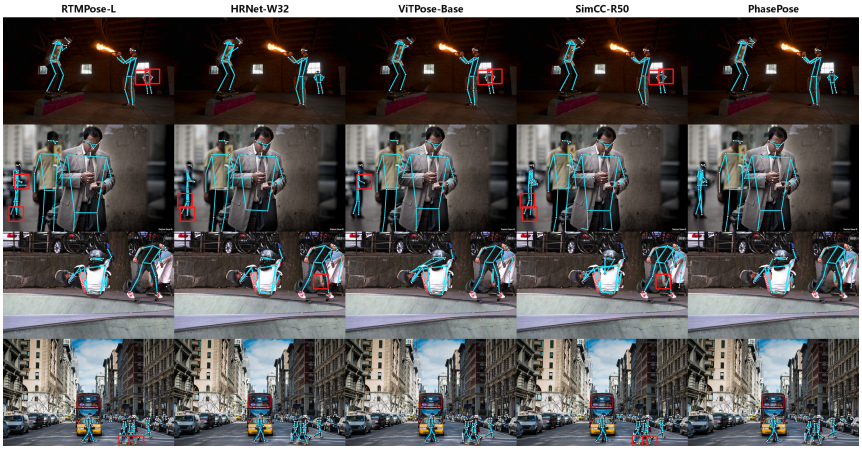

Figure 4: COCO val2017 comparison under identical MMPose Faster R-CNN detections; red boxes mark distal-joint errors corrected by PhasePose.

4.4 Ablation

Table 4(a) isolates the functional blocks. Recurrent correction is the largest single contributor (-0.96 AP), followed by phase-detail memory (-0.77) and the spectral distribution state (-0.74). Removing adaptive analysis costs 0.54 AP, while removing recurrent conditioning costs 0.41 AP. These drops show that the gain is distributed across preserved phase evidence and uncertainty-aware iterative correction rather than a single enlarged head.

Table 4: COCO val2017 ablations at 256×192. ΔAP is relative to the complete model.

(a) Functional blocks

Variant	AP	ΔAP
Full PhasePose	**76.70**	–
w/o recurrent correction	75.74	-0.96
w/o phase detail memory	75.93	-0.77
w/o spectral distribution state	75.96	-0.74
w/o adaptive analysis	76.16	-0.54
w/o recurrent conditioning	76.29	-0.41

(b) Recurrent depth

Steps	AP	ΔAP
0	75.74	-0.96
1	76.09	-0.61
2	76.48	-0.22
3 (Full)	**76.70**	–

(c) Phase representation

Variant	AP	ΔAP
No detail memory	75.93	-0.77
Raw phase	76.04	-0.66
Centered phase	76.19	-0.51
Fixed Haar	76.35	-0.35
Learned phase mixing	76.48	-0.22
Haar + $\mathrm{SO}(3)$ **(Full)**	**76.70**	–

Depth. Table 4(b) separates recurrent reuse from model size. Starting from the 75.74 AP zero-step prediction, one and two shared corrections reach 76.09 and 76.48 AP, and the third reaches 76.70. Because the corrector weights are shared, this progression adds iterative state evolution without multiplying corrector parameters.

Phase representation. Table 4(c) separates the benefit of retaining directional detail from the choice of phase basis. Removing detail drops AP to 75.93; raw and centered phases recover to 76.04 and 76.19, while fixed Haar reaches 76.35. Generic learned mixing improves to 76.48, and the Haar-anchored grouped $\mathrm{SO}(3)$ rotation reaches 76.70, adding 0.35 AP over fixed Haar and 0.22 AP over unconstrained mixing.

Routing and supervision. Entropy-conditioned memory routing reaches 76.70 AP versus 76.39/76.49 for uniform/static routing. Removing the learned recurrence identifier gives 76.29 AP. KL supervision alone reaches 76.19 AP; intermediate KL and final OKS provide complementary gains, with the complete objective returning 76.70. Additional detailed controls are reported in the supplement.

4.5 Limitations

PhasePose cannot recover evidence already discarded before the stride-16 phase source, so extremely small or occluded distal landmarks remain difficult. Detector crop errors can therefore propagate into the pose stage, especially when a limb is truncated or two people overlap. The fixed within-person skeleton also does not explicitly model cross-person contact in crowded scenes. FLOPs do not fully predict eager GPU latency because the recurrent corrector contains several small kernels. Additional failure cases and deployment analysis are provided in the supplement.

5 Conclusion

PhasePose separates coordinate precision from feature-map density by preserving stride-16 sampling phases and recurrently refining compact spectral summaries of SimCC distributions on a stride-32 lattice. Haar-anchored $\mathrm{SO}(3)$ detail routing preserves an analytic low-frequency path while adapting the directional subspace; DCT-anchored low-rank analysis and entropy-conditioned memory reading make the recurrent state adaptive without abandoning its spectral prior. Across normal and low resolutions and four pose topologies, phase memory, distribution-aware state, and skeletal exchange provide complementary gains without high-resolution reconstruction. This shifts efficiency from spatial reconstruction toward compact, uncertainty-aware correction. More broadly, precise localization can remain compatible with a coarse spatial workspace when phase evidence and distribution shape are preserved explicitly. The results also indicate where the low-resolution assumption should be tested next: detector uncertainty, cross-person interactions, and temporally linked evidence could be incorporated into the same compact memory-and-correction framework without reverting to dense spatial restoration. These extensions are left for future work.

AI USE STATEMENT

Generative AI tools were used to assist with LaTeX formatting and manuscript presentation, including adapting the source to the ICLR 2027 template and checking table layout. All AI-assisted edits were reviewed by the authors. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The architecture and training objective are specified in Sections 3–4, while the supplementary material provides the full optimization schedule, data augmentation, dataset-specific skeleton graphs, controlled ablations, and latency protocol. These details are intended to make the reported configurations and evaluation procedures reproducible.

REFERENCES

X. An, L. Zhao, C. Gong, N. Wang, D. Wang, and J. Yang. Sharpose: Sparse high-resolution representation for human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 691–699, 2024.

M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

Yuanhao Cai, Zhijian Wang, Zhengxiong Luo, B. Yin, A. Du, H. Wang, et al. Learning delicate local representations for multi-person pose estimation. In *European Conference on Computer Vision*, pp. 455–472. Springer, 2020.

H. Chen, R. Feng, S. Wu, H. Xu, F. Zhou, and Z. Liu. 2d human pose estimation: A survey. *Multimedia Systems*, 29(5):3115–3138, 2023.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

S. Hu, H. Sun, D. Wei, X. Sun, and J. Wang. Continuous heatmap regression for pose estimation via implicit neural representation. *Advances in Neural Information Processing Systems*, 37: 102036–102055, 2024.

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*, pp. 5700–5709, 2020.

S. Jiang, D. Zhang, R. Yan, et al. Eliminating semantic ambiguity in human pose estimation via stable feature upsampling. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024.

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. In *European Conference on Computer Vision*, pp. 196–214. Springer, 2020.

Cansu Korkmaz, A. Murat Tekalp, and Zafer Dogan. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5926–5936, 2024.

Bohao Li, Zhicheng Cao, Huixian Li, and Yangming Guo. Cigpose: Causal intervention graph neural network for whole-body pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23484–23494, 2026.

J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10863–10872, 2019.

J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11025–11034, 2021a.

Siyuan Li, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, and Stan Z. Li. Moganet: Multi-order gated aggregation network. In *International Conference on Learning Representations*, 2024.

Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S. Xia, and E. Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11313–11322, 2021b.

Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, and S. Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pp. 89–106. Springer, 2022.

T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1491–1500, 2024.

Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022.

Stéphane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pp. 483–499. Springer, 2016.

M. Purkrabek and J. Matas. Probpose: A probabilistic approach to 2d human pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27124–27133, 2025.

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*, pp. 5693–5703, 2019.

Pengzhan Sun, Kerui Gu, Yunsong Wang, Linlin Yang, and Angela Yao. Rethinking visibility in human pose estimation: Occluded pose reasoning via transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5903–5912, 2024.

D. Wang and S. Zhang. Spatial-aware regression for keypoint localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 624–633, 2024.

H. Wang, J. Liu, J. Tang, and G. Wu. Lightweight super-resolution head for human pose estimation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2353–2361, 2023.

H. Wang, J. Liu, J. Tang, G. Wu, B. Xu, Y. Chou, and Y. Wang. Gtpt: Group-based token pruning transformer for efficient human pose estimation. In *European Conference on Computer Vision*, pp. 213–230. Springer, 2024.

Y. Wang, M. Li, H. Cai, W. Chen, and S. Han. Lite pose: Efficient architecture design for 2d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13126–13136, 2022.

B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 466–481, 2018.

Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: Searching for whole-body human pose estimation in the wild. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022a.

Y. Xu, J. Zhang, Q. Zhang, and D. Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022b.

Y. Xu, J. Zhang, Q. Zhang, and D. Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023.

Yalong Xu, Lin Zhao, Chen Gong, Guangyu Li, Di Wang, and Nannan Wang. Dynpose: Largely improving the efficiency of human pose estimation by a simple dynamic framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1160–1169, 2025.

S. Yang, Z. Quan, M. Nie, and W. Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11802–11812, 2021.

Z. Yang, A. Zeng, C. Yuan, and Y. Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4210–4220, 2023.

Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34: 7281–7293, 2021.

Wang Zeng, Sheng Jin, Lumin Xu, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Tcformer: Visual recognition via token clustering transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7093–7102, 2020.