Learning Topology-Aware Dynamic Associations for Robust Multi-Person Pose Estimation

Shengnan Hu*, **Yandong Liu***, **Jiangnan Liu, Yahong Chen**†

Central China Normal University, Wuhan, China
{shengnanhu@, lyd1013@mails., jiangnanliu@mails., chenyahong123@}ccnu.edu.cn

Abstract

Multi-person pose estimation in real-world scenarios remains a challenging task due to frequent occlusions, scale variations, and complex human interactions. Existing methods often rely on fixed keypoint association patterns that fail to capture the dynamic and context-dependent nature of human body topologies, leading to misalignment and false detections. In this work, we propose a topology-aware dynamic association framework that adaptively models inter-keypoint relationships conditioned on local context and pose topology. The proposed framework consists of three stages: a human-to-keypoint detection module for coarse localization, a dynamic keypoint association module that learns flexible connectivity patterns between joints, and a fine-grained refinement module for precise pose adjustment. By integrating topological priors into dynamic learning and multi-stage optimization, our proposed method effectively mitigates the issues caused by occlusions and overlapping instances. Extensive experiments on benchmark datasets demonstrate that our approach achieves state-of-the-art performance, especially in crowded and occlusion-heavy scenes.

Code — https://github.com/suiyingliuxin/TopoDA

Introduction

Multi-person pose estimation (MPPE) aims to detect the body keypoints of all individuals in an image and is a fundamental task in computer vision, supporting applications such as human–computer interaction, behavior understanding, and sports analytics (Sun et al. 2019; Li et al. 2021). While significant progress has been made in recent years, achieving accurate and robust keypoint localization in real-world scenes remains challenging, especially under conditions of occlusion, scale variation, and dense human interactions.

Existing approaches can be broadly categorized into top-down, bottom-up, and single-stage frameworks (Mao et al. 2021). Top-down methods localize each person via a detector and apply pose estimation within fixed bounding boxes. Although effective in uncluttered environments, these methods suffer from cascading detection errors and become computationally inefficient in crowded scenes. Bottom-up methods detect all keypoints first and group them into individuals, but they often produce incorrect associations when people overlap or interact. Single-stage models aim to unify detection and estimation in a shared pipeline, offering better efficiency, yet they typically rely on fixed receptive fields and static keypoint association priors, which limits their generalization to complex poses.

A fundamental limitation across these paradigms is the reliance on predefined and rigid keypoint topologies—assuming that all human poses conform to a fixed skeletal structure. However, human poses are highly dynamic: their joint connectivity patterns vary dramatically with occlusion, body orientation, and interaction context. Such static assumptions are inadequate for resolving keypoint ambiguity, often leading to misaligned joints, missing limbs, or incorrect merging of parts from adjacent individuals.

To address these limitations, we propose **TopoDA**, a novel Topology-Aware Dynamic Association framework that estimates multi-person poses through hierarchical and adaptive reasoning. TopoDA is built upon a coarse-to-fine pipeline with three progressively refined stages: (1) a Human-to-Keypoint Detection module with Locally-Aware Feature Alignment(LAFA) that resolves multi-scale spatial inconsistencies, improving keypoint localization for small or occluded persons; (2) a Dynamic Keypoint Box Modulation module that adaptively adjusts keypoint receptive fields based on both global human context and local semantic cues, enhancing spatial flexibility; and (3) a Topology-Aware Keypoint Refinement module with grouped attention that captures dynamic inter-keypoint dependencies without relying on fixed skeleton priors, enabling the model to reason over flexible, scene-conditioned topologies.

Through this coarse-to-fine design, **TopoDA** integrates adaptive region modeling, semantic-guided feature fusion, and implicit topology reasoning into a unified and scalable pipeline. It achieves strong robustness in highly interactive scenes and delivers consistent performance under occlusion and scale variation.

In summary, our main contributions are:

- We propose TopoDA, a topology-aware dynamic as-

*These authors contributed equally.
†Corresponding author.

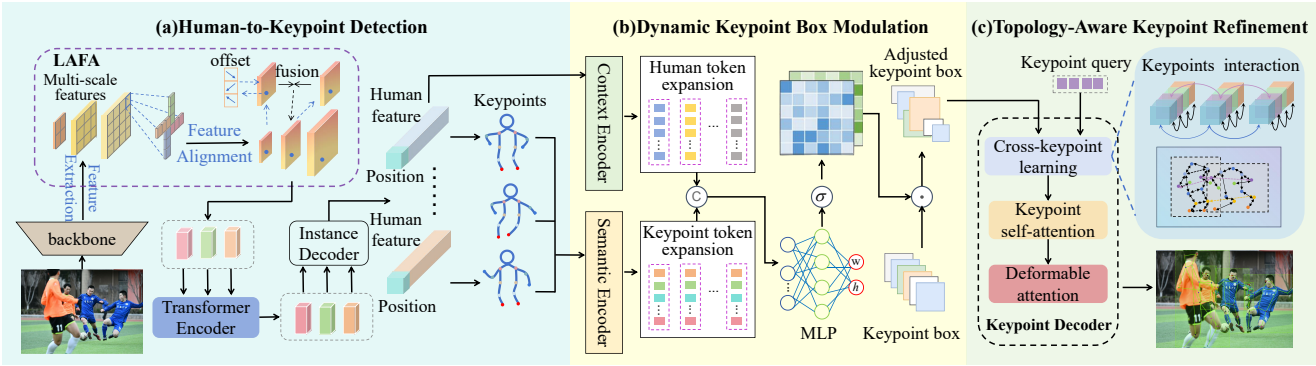

Figure 1: Overview of the proposed **TopoDA** architecture. (a) Human-to-keypoint detection with locally-aware feature alignment (LAFA). (b) Dynamic keypoint box modulation for adaptive receptive field adjustment. (c) Topology-aware keypoint refinement for fine-grained pose estimation.

sociation framework that progressively refines multi-person pose estimation through a hierarchical and adaptive pipeline.

- We design LAFA, a locally-aware multi-scale feature alignment module that alleviates spatial misalignment in early-stage keypoint localization.
- We develop a dynamic keypoint box modulation mechanism that adjusts receptive fields contextually to improve detection under scale and occlusion variation.
- We propose a cross-keypoint learning module that models inter-joint dependencies via grouped attention, enabling implicit and adaptable topology reasoning.

Related Work

Multi-person pose estimation methods can be broadly categorized into top-down, bottom-up, and single-stage approaches.

Top-down Methods. Top-down approaches (Xiao, Wu, and Wei 2018; Li et al. 2021; Sun et al. 2019; Ding et al. 2022) first detect human instances and then estimate keypoints within each bounding box. While effective in simple scenes, their performance degrades in crowded or occluded settings due to dependence on detection quality. Representative works include Mask R-CNN (He et al. 2017), which integrates RoI-based feature extraction, and HRFormer (Yuan et al. 2021), which leverages transformer-based backbones. Although efficient variants like RTMPose (Jiang et al. 2023) reduce inference cost, these methods still suffer from scalability issues as their complexity grows linearly with the number of people.

Bottom-up Methods. Bottom-up methods detect all keypoints first and then group them into person instances via part association or embedding strategies. Early regression-based methods (Toshev and Szegedy 2014) were limited by their lack of spatial granularity. Heatmap-based pipelines such as CPM (Wei et al. 2016) and HigherHRNet (Cheng et al. 2020) improved accuracy on small or overlapping targets. Grouping techniques include graph optimization (Pishchulin et al. 2016; Insafutdinov et al. 2016), part affinity fields (Cao et al. 2017), and associative embeddings (Newell, Huang, and Deng 2017). Although bottom-up approaches are typically faster, they often struggle with ambiguous grouping in cluttered scenes.

Single-stage Methods. Single-stage frameworks (Mao et al. 2021; Lu et al. 2024) unify person detection and keypoint estimation in an end-to-end architecture, offering better efficiency and stronger contextual modeling. FCPose (Mao et al. 2021) and CID (Wang and Zhang 2022) proposed instance-aware feature decoupling to reduce inter-person interference. ED-Pose (Yang et al. 2023) introduced cascaded detection decoders for explicit global-local learning. GroupPose (Liu et al. 2023) further simplified the pipeline via group attention across instance and keypoint queries. While recent single-stage models have improved overall efficiency and reduced reliance on hand-crafted post-processing, they still depend on fixed keypoint association patterns and static spatial priors, which limit their adaptability in challenging scenes involving heavy occlusions, dense interactions, or diverse human poses. These limitations highlight the need for a framework that can adaptively model keypoint relationships and spatial regions based on context.

Method

Overview

In this section, we introduce TopoDA, a topology-aware dynamic association framework for robust multi-person pose estimation in complex scenes. As illustrated in Fig.1, TopoDA adopts a three-stage coarse-to-fine architecture that progressively refines pose predictions, reflecting the hierarchical nature of visual perception—from global instance-level understanding to fine-grained keypoint localization. This design is motivated by two key observations: first, modeling human-level context before keypoint inference improves robustness and reduces error propagation; second, keypoint associations are inherently dynamic and context-

dependent, and thus benefit from iterative refinement across representation levels. The framework begins with a joint human-to-keypoint detection stage, which estimates person instances and their corresponding keypoints in a unified manner, mitigating the decoupling issues commonly observed in traditional top-down pipelines. Building upon this initial prediction, the second stage introduces dynamic keypoint box modulation, where the spatial extent of each keypoint region is adaptively adjusted based on both global context and local semantic cues, enhancing the model's ability to handle scale variation and peripheral joints. Finally, the fine pose adjustment stage incorporates a cross-keypoint learning module that captures both intra- and inter-person dependencies through grouped attention, enabling the model to reason over flexible topological relationships and refine keypoint assignments under challenging conditions such as occlusion and crowding.

Human-to-Keypoint Detection with Locally-Aware Feature Alignment

Multi-scale feature fusion is essential for detecting persons at different scales, yet naive fusion (Lin et al. 2017; Huang et al. 2021) often introduces spatial misalignment and degrades fine-grained details. To address this limitation, we propose locally-aware feature alignment (LAFA) module that simultaneously preserves spatial precision and insufficient preservation of fine-grained local details essential for accurate keypoint detection in occluded scenarios, which is critical for robust pose estimation under occlusions and scale variations.

As illustrated in Fig. 2, LAFA consists of two key components: feature extraction and feature alignment. The feature extraction branch employs a Pinwheel-shaped Convolution, which is effective at capturing fine-grained local patterns and improving the detection of small-scale human instances. In parallel, the feature alignment branch aligns features across resolutions, ensuring spatial coherence between semantic and detailed representations. By jointly modeling local sensitivity and spatial correspondence, LAFA enables the network to retain fine-grained cues while leveraging multi-scale context—an essential capability for robust pose estimation under occlusion and scale variation.

Feature Extraction. Given a set of multi-scale feature maps extracted from the backbone, denoted as $\{F_1, F_2, ..., F_n\}$, where $F_i \in \mathbb{R}^{C\times H_i\times W_i}$ represents the feature map at scale i, we introduce a directional feature extraction mechanism to enhance local structural representation. Specifically, each F_i is first divided into four groups along the channel dimension, and each group is asymmetrically padded in one of the four cardinal directions (up, down, left, or right) to expand its receptive field while preserving spatial alignment.

To capture elongated patterns and spatial dependencies along the principal axes of human poses, we apply directional convolutions with kernel shapes $(1 \times k)$ and $(k \times 1)$ to each padded group. The resulting directional features $\{F_i^{\text{up}}, F_i^{\text{down}}, F_i^{\text{left}}, F_i^{\text{right}}\}$ are then aggregated via channel-wise concatenation and fused using a 2×2 convolution to

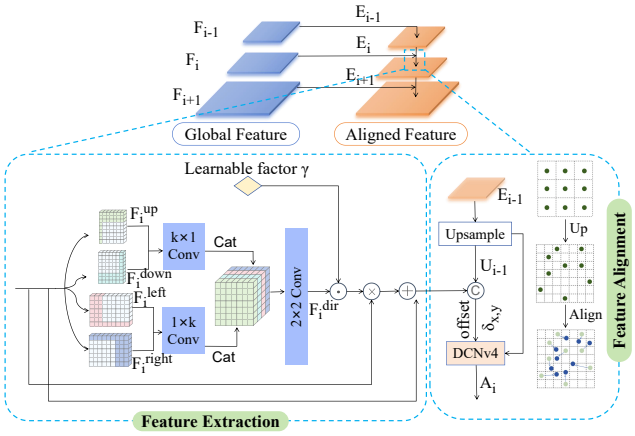

Figure 2: Locally-aware feature alignment. Perform feature extraction and feature alignment on multi-scale feature maps, and adopt Pinwheel-shaped Convolution to enhance local perception of human bodies.

produce a unified representation:

$$F_i^{\text{dir}} = \text{Conv}_{2\times 2}\left(\text{Concat}(F_i^{\text{up}}, F_i^{\text{down}}, F_i^{\text{left}}, F_i^{\text{right}})\right). \quad (1)$$

To adaptively emphasize the most informative spatial directions, we introduce a learnable modulation factor $\gamma \in \mathbb{R}^{C\times H_i\times W_i}$, which dynamically reweights the directional responses:

$$F_i^{\text{aligned}} = \gamma \odot F_i^{\text{dir}}, \quad (2)$$

where $\odot$ denotes element-wise multiplication. This mechanism allows the model to selectively enhance relevant structural cues based on the contextual demands of each scene.

Feature Alignment. To ensure precise spatial correspondence between multi-scale features, we incorporate a deformable alignment strategy that refines the resolution transition within the LAFA module. Given a high-resolution feature map $\mathbf{E}_i \in \mathbb{R}^{C\times H_i\times W_i}$ and its corresponding lower-resolution counterpart $\mathbf{E}_{i-1} \in \mathbb{R}^{C\times \frac{H_i}{2}\times \frac{W_i}{2}}$, we first upsample the lower-resolution features to match the target spatial size:

$$\mathbf{U}_{i-1} = \text{Upsample}_{\times 2}(\mathbf{E}_{i-1}), \quad (3)$$

where $\text{Upsample}_{\times 2}$ denotes bilinear or nearest-neighbor interpolation.

To correct for spatial misalignment introduced by scale variation and convolutional downsampling, we predict a dense offset field $\delta_{x,y}$ using Deformable Convolution v4 (DCNv4), conditioned on both the upsampled feature $\mathbf{U}_{i-1}$ and the high-resolution reference $\mathbf{E}_i$:

$$\delta_{x,y} = \text{DCNv4}(\mathbf{U}_{i-1}, \mathbf{E}_i), \quad (4)$$

where $\delta_{x,y} \in \mathbb{R}^{2\times H_i\times W_i}$ encodes pixel-wise 2D offsets.

The aligned feature map $\mathbf{A}_i$ is then computed via deformable sampling:

$$\mathbf{A}_i = \sum_{p\in\mathcal{R}} w(p)\, \mathbf{U}_{i-1}(p_0 + p + \delta_{x,y}), \quad (5)$$

where $\mathcal{R}$ defines the sampling region, p_0 denotes the reference location, $\delta_{x,y}$ is the predicted offset, and $w(p)$ represents bilinear interpolation weights.

This adaptive alignment ensures spatial consistency between resolutions, enabling effective fusion of semantic and structural cues across scales.

Instance to Keypoint. To bridge global human-level semantics with local keypoint-level representations, we introduce a human-to-keypoint detection stage that explicitly models topological associations between instances and their joints. As illustrated in Fig. 1(a), the multi-scale features processed by the LAFA module are flattened and passed through a transformer encoder to enable information exchange across resolutions. A set of learnable instance queries then attend to these encoded features via self- and cross-attention, yielding contextualized human-level embeddings and their corresponding position queries.

Each instance embedding is further expanded into a set of keypoint queries, ensuring that the local pose reasoning for each person is grounded in the corresponding global context. For initialization, we decompose keypoint position into center coordinates and spatial extent. Specifically, the center (x_k, y_k) of each keypoint is directly predicted from its associated instance embedding using a lightweight regression head:

$$(x_k^n, y_k^n) = \text{FFN}_{\text{center}}(f_h^n), \quad k = 1, \ldots, K, \tag{6}$$

where f_h^n is the n-th human instance embedding and K denotes the number of keypoints.

The keypoint region size (w_k, h_k) is initialized from human prior heuristics and adaptively refined in the next stage via our dynamic keypoint box modulation. This structured initialization provides robust topological priors that guide downstream refinement and improve resilience under occlusion and interaction.

Dynamic Keypoint Box Modulation

The success of keypoint detection heavily relies on the appropriate receptive field size for each keypoint query. However, existing methods (Liu et al. 2023; Mao et al. 2021) typically employ fixed-size detection boxes, which suffer from several fundamental limitations. Oversized boxes introduce background noise and increase the probability of incorrect detection. Conversely, undersized boxes fail to capture sufficient contextual information for keypoint disambiguation, potentially leading to misclassification between semantically similar keypoints from different individuals. Moreover, the optimal box size varies significantly across different keypoint types and human scales, making static approaches inherently suboptimal for robust multi-person pose estimation. As shown in Fig.1 (b),We propose a novel dynamic keypoint box modulation approach that adaptively adjusts keypoint box sizes by jointly modeling human context features and keypoint semantics.

Human Context Encoder. Accurate keypoint localization requires a comprehensive understanding of the contextual information surrounding each human instance. The human instance queries $Q_h \in \mathbb{R}^{N\times C}$ from the previous stage are taken as input, where N represents the number of detected individuals and C denotes the feature dimension. We transform these features into a more informative contextual representation through:

$$Q_{he} = f_1(Q_h) \in \mathbb{R}^{N\times D}. \tag{7}$$

Keypoint Semantic Encoder. Keypoint features contain rich semantic information that reflects the characteristics of different body parts. To effectively capture these semantics, we transform the initial keypoint queries $Q_k \in \mathbb{R}^{K\times C}$, where K represents the number of keypoints and C is the feature dimension, into a refined semantic representation. This transformation is performed using a learnable function:

$$Q_{ke} = f_2(Q_k) \in \mathbb{R}^{K\times D}. \tag{8}$$

Adaptive Box Size Prediction. To capture the interdependence between keypoints and human instances, the encoded features Q_{ke} and Q_{he} are fused. First, Q_{ke} is expanded to match the human instance dimensions, and Q_{he} is expanded to match the keypoint dimensions:

$$\mathbf{Q}_{he}^{'} \in \mathbb{R}^{N\times K\times D} = \text{expand}(\mathbf{Q}_{he}, K), \tag{9}$$

$$\mathbf{Q}_{ke}^{'} \in \mathbb{R}^{N\times K\times D} = \text{expand}(\mathbf{Q}_{ke}, N). \tag{10}$$

The expanded features are concatenated along the feature dimension to form a combined representation:

$$Q_{\text{hk}} \in \mathbb{R}^{N\times K\times 2D} = \text{concat}(Q_{he}^{'}, Q_{ke}^{'}). \tag{11}$$

This fused representation is then processed by MLP and sigmoid to generate adaptive scaling factors:

$$\alpha \in \mathbb{R}^{N\times K\times 2} = \sigma(MLP(Q_{\text{hk}})). \tag{12}$$

These scaling factors $\alpha \in \mathbb{R}^{N\times K\times 2}$ are then multiplied with the initial keypoint boxes $K_b \in \mathbb{R}^{N\times K\times 2}$ to dynamically adjust the width and height of each keypoint box.

Topology-Aware Keypoint Refinement

The final stage of the proposed TopoDA focuses on fine-grained pose refinement through topology-aware reasoning. While earlier stages provide coarse human detection and adaptive keypoint localization, accurate pose estimation in complex scenes requires explicit modeling of both intra-person joint structures and inter-person keypoint associations, especially in crowded scenarios with overlapping or ambiguous keypoints.

As illustrated in Fig.1 (c), this stage operates on the dynamically adjusted keypoint features from the previous stage and employs a multi-component architecture to capture different aspects of keypoint dependencies. Cross-keypoint learning models both intra-person skeletal topology and inter-person semantic consistency across the same keypoint types. Keypoint self-attention enhances individual keypoint representations through contextual reasoning. Deformable attention enables adaptive spatial feature aggregation for precise localization.

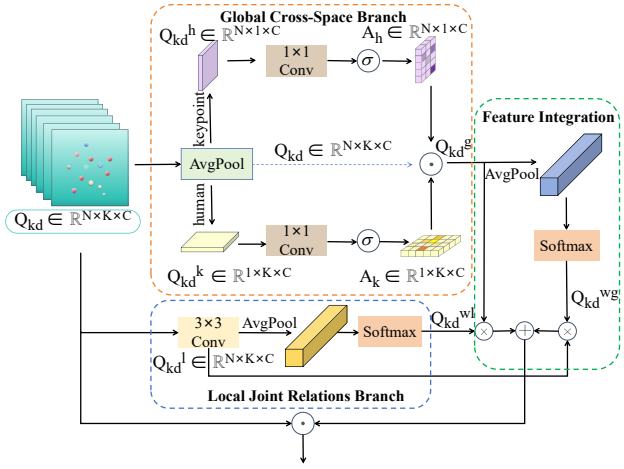

Figure 3: Cross-keypoint learning. It simultaneously learns the spatial correlations between keypoints within human bodies and the semantic consistency across instances of the same keypoint type.

Cross-keypoint Learning. The core of our refinement module is the cross-keypoint learning mechanism, which explicitly models spatial dependencies across human instances and keypoints. To construct comprehensive keypoint features $Q_{kd} \in \mathbb{R}^{N\times K\times C}$, where N is the number of persons and K the number of keypoints, we fuse spatial information from adjusted keypoint boxes with semantic representations from keypoint queries. As shown in Fig. 3, we adopt a dual-space relational modeling strategy to capture both global and local dependencies.

To capture global dependencies across keypoints and human instances, we perform global average pooling along the keypoint and human dimensions, respectively:

$$Q^h_{kd} \in \mathbb{R}^{N\times 1\times C} = \text{GAP}_{keypoint}(Q_{kd}), \quad (13)$$

$$Q^k_{kd} \in \mathbb{R}^{1\times K\times C} = \text{GAP}_{human}(Q_{kd}). \quad (14)$$

These aggregated features are then passed through 1×1 convolutions followed by sigmoid activations to generate attention weights A_h and A_k. By applying these weights to the original keypoint features, we recalibrate the global responses:

$$Q^g_{kd} = Q_{kd} \odot A_h \odot A_k. \quad (15)$$

This allows the model to reason over inter-person correspondences and semantic relationships among keypoints of the same type.

In parallel, we enhance local joint relations by applying 3×3 convolutions to Q_{kd}, producing fine-grained spatial features Q^l_{kd}. This operation preserves local structural patterns, which are essential for precise joint localization, especially in crowded or occluded scenarios.

Feature Integration. We apply global average pooling to the cross-space feature Q^g_{kd} to generate a descriptor $Q^{wg}_{kd} \in \mathbb{R}^{1\times 1\times C}$ encoding keypoint interactions across instances. After Softmax normalization, this descriptor is multiplied with Q^l_{kd}, producing a spatially-sensitive attention representation $Q^{f1}_{kd} \in \mathbb{R}^{N\times K\times 1}$. We construct a complementary attention representation $Q^{f2}_{kd} \in \mathbb{R}^{N\times K\times 1}$ capturing different interaction patterns. These representations are fused through element-wise addition, forming a comprehensive keypoint interaction-aware mechanism that improves localization of occluded keypoints in crowded scenes.

Experiments

Experimental Settings

Datasets. We conducted experiments on two representative human pose estimation datasets: COCO (Lin et al. 2014) and CrowdPose (Li et al. 2019), to evaluate the model's performance in both general and crowded scenarios. **COCO** is the most widely used benchmark dataset in human pose estimation, containing over 200k images and 250k annotated human instances, with 17 keypoints annotated for each instance. **CrowdPose** is a dataset specifically focused on crowded scenarios, containing 20k images and 80,000 annotated instances, with 14 keypoints annotated for each instance. This dataset is specifically designed with three levels of crowding - low, medium, and high density, facilitating the analysis of model performance variations across different crowd density scenarios.

Implementation Details. Following prior works (Liu et al. 2023; Yang et al. 2023), we adopt Swin-T pretrained on ImageNet as the backbone. Standard data augmentation techniques, including random flipping, cropping, and resizing, are applied during training. The model is optimized with AdamW (Kingma 2014; Loshchilov 2017), using a base learning rate of 1×10^{-4} and 1×10^{-5} for backbone layers. During inference, images are resized with the shorter side set to 800 pixels and the longer side capped at 1333 pixels, maintaining aspect ratio. For quantitative assessment, we follow the standard evaluation metric (Zheng et al. 2023) and adopt the Object Keypoint Similarity (OKS) as our evaluation protocol.

Comparison with State-of-the-art Methods

Results on CrowdPose. Our method is primarily designed for complex scenarios involving severe occlusions and dense crowds, therefore we first conduct evaluations on the CrowdPose testing set, which is specifically constructed for complex scenarios. As shown in Table 1, our method achieves state-of-the-art performance with an AP of 73.5, demonstrating superior performance in scenarios containing severe occlusions and overlapping poses. These results further confirm the effectiveness of our framework in complex real-world environments, where accurate pose estimation remains highly challenging.

Results on COCO. To further validate the generalization capability of our method, we evaluate it on the COCO val2017 dataset, which contains relatively simple scenarios. As shown in Table 2, our method achieves an AP of 73.8, maintaining competitive results. This demonstrates that our

Method	AP↑	AP^{50}↑	AP^{75}↑	AP^{E}↑	AP^{M}↑	AP^{H}↑
Top-down methods						
Mask-R-CNN (He et al. 2017)	57.2	83.5	60.3	69.4	57.9	45.8
AlphaPose (Fang et al. 2022)	61.0	81.3	66.0	71.2	61.4	51.1
SPPE (Li et al. 2019)	66.0	84.2	71.5	75.5	66.3	57.4
SDPose-S-V1 (Chen et al. 2024)	57.3	-	-	-	-	-
Bottom-up methods						
OpenPose (Cao et al. 2017)	-	-	-	62.7	48.7	32.3
HigherHRNet† (Cheng et al. 2020)	65.9	86.4	70.6	73.3	66.5	57.9
DEKR† (Geng et al. 2021)	67.3	86.4	72.2	74.6	68.1	58.7
Single-stage methods						
PINet (Wang, Zhang, and Hua 2021)	68.9	88.7	74.7	75.4	69.6	61.5
ED-Pose(ResNet-50) (Yang et al. 2023)	69.9	88.6	75.8	77.7	70.6	60.9
RTMO-s (Lu et al. 2024)	67.3	-	-	-	-	-
LMFormer (Li, Tang, and Li 2024)	62.6	80.9	67.6	-	-	-
Ours						
TopoDA(Resnet-50)	71.4	89.6	77.5	79.3	72.1	62.3
TopoDA(Swin-T)	**73.5**	**90.4**	**80.1**	**81.1**	**74.2**	**64.5**

Table 1: **Comparison with state-of-the-art methods on CrowdPose testing set.**† indicates that flip test augmentation is used.

Method	Backbone	AP↑	AP^{50}↑	AP^{75}↑	AP^{M}↑	AP^{L}↑
Top-down methods						
Mask-R-CNN (He et al. 2017)	ResNet-50	65.5	87.2	71.1	61.3	73.4
PRTR† (Li et al. 2021)	ResNet-50	68.2	88.2	75.2	63.2	76.2
SDPose-T (Chen et al. 2024)	HRNet-w32	69.7	88.1	77.3	66.1	76.6
Bottom-up methods						
HigherHRNet† (Cheng et al. 2020)	HRNet-w32	67.1	86.2	73.0	61.5	76.1
DEKR† (Geng et al. 2021)	HRNet-w32	68.0	86.7	74.5	62.1	77.7
LOGO-CAP† (Xue et al. 2022)	HRNet-w32	69.6	87.5	75.9	64.1	78.0
Single-stage methods						
FCPose (Mao et al. 2021)	ResNet-50	63.0	85.9	68.9	59.1	70.3
PETR (Shi et al. 2022)	ResNet-50	68.8	87.5	76.3	62.7	77.7
ED-Pose (Yang et al. 2023)	ResNet-50	71.6	89.6	78.1	65.9	79.8
LMFormer (Li, Tang, and Li 2024)	LMFormer-L	70.5	88.4	77.6	-	-
Ours						
TopoDA	ResNet-50	71.9	89.8	78.7	66.2	79.8
TopoDA	Swin-T	**73.8**	**90. 6**	**80.9**	**67.9**	**81.8**

Table 2: **Comparisons with state-of-the-art methods on COCO val2017.** † indicates that flip test augmentation is used.

method not only excels in complex scenarios but also maintains strong performance in regular scenarios, fully validating the good generalization capability and broad applicability of the proposed approach.

Ablation Study

We perform ablation experiments on CrowdPose to evaluate the effectiveness of each module in TopoDA, as shown in Table 3. It could be observed that dynamic keypoint box modulation (DKBM) improves localization precision by adaptively adjusting receptive fields, achieving a 1.1 AP gain over the baseline. Cross-keypoint learning (CL) further boosts AP by 1.3 by modeling spatial relationships within and across human instances. Locally-aware feature alignment (LAFA) provides the largest single-module improvement (+1.6 AP), highlighting its importance in mitigating multi-scale misalignment, especially under occlusions.

Furthermore, combining DKBM and CL yields additional gains, demonstrating their complementary roles. Integrating all three modules achieves the best performance, with 71.4 AP, 89.6 AP^{50}, and 77.5 AP^{75}. These results confirm that each component addresses a specific challenge, and their combination leads to significant improvements in pose estimation accuracy under complex real-world conditions.

DKBM	CL	LAFA	AP↑	AP^{50}↑	AP^{75}↑
			68.7	88.0	74.9
✓			69.8	88.8	75.6
	✓		70.0	88.6	76.0
		✓	70.3	88.9	76.4
✓	✓		70.4	89.3	76.1
✓	✓	✓	**71.4**	**89.6**	**77.5**

Table 3: Ablation study of the components on CrowdPose.

Study on Feature Alignment Strategies. To investigate the role of feature alignment in multi-task pose estimation, we conduct ablation studies comparing four alignment strategies, as summarized in Table 4: (1) DF (Zhang et al. 2021): direct multi-scale fusion without alignment; (2) ADD (Farias and Maziero 2023): element-wise addition of predicted offsets; (3) ConvOL (Zhong et al. 2022): offset learning via standard convolution; and (4) DCNv4OL: alignment using deformable convolution for adaptive sampling.

Among them, DF yields the lowest performance, underscoring the importance of effective alignment. ADD provides minor improvements but fails to resolve semantic

Strategy	$AP\uparrow$	$AP^{50}\uparrow$	$AP^{75}\uparrow$
DF (Zhang et al. 2021)	70.7	89.2	76.8
ADD (Farias and Maziero 2023)	70.9	89.3	77.1
ConvOL (Zhong et al. 2022)	71.1	89.4	77.2
DCNv4OL (Ours)	**71.4**	**89.6**	**77.5**

Table 4: Comparison of different feature alignment strategies.

Method	Params	$AP\uparrow$	$AP^{50}\uparrow$	$AP^{75}\uparrow$
FPN (Lin et al. 2017)	50.6G	70.1	88.3	76.1
PAFPN (Liu et al. 2018)	54.1G	70.5	88.6	76.2
BiFPN (Tan, Pang, and Le 2020)	48.7G	70.8	88.3	76.5
LAFA (Ours)	53.9G	**71.4**	**89.6**	**77.5**

Table 5: Comparison of different FPN architectures.

and spatial inconsistencies. ConvOL enhances alignment via learned offsets, yet is constrained by fixed sampling patterns. DCNv4OL outperforms all baselines by dynamically focusing on informative regions, achieving consistent gains across AP, AP^{50}, and AP^{75}. These results demonstrate the advantage of deformable alignment for precise and flexible multi-scale feature integration.

Study on FPN Architectures. To evaluate the effectiveness of our proposed LAFA module within different feature pyramid designs, we compare it against several widely used FPN varians. As shown in Table 5, LAFA introduces a reasonable parameter overhead of 53.9G compared to FPN (Lin et al. 2017)'s 50.6G, representing a modest 6.5% increase. Notably, LAFA achieves better parameter efficiency than PAFPN (Liu et al. 2018) while delivering significantly superior performance. Although BiFPN (Tan, Pang, and Le 2020) requires fewer parameters (48.7G), LAFA's +0.6 AP performance advantage justifies the additional computational cost, particularly for accuracy-critical applications involving complex occlusion scenarios. These results confirm the advantage of LAFA in scenarios where precise localization is critical.

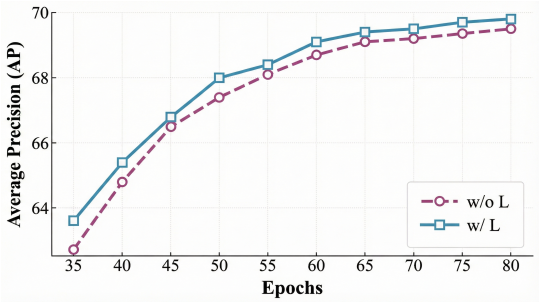

Figure 4: The results with/without dynamic keypoint box constraint.

Study on Dynamic Keypoint Box Modulation. To assess the effectiveness of the proposed dynamic box constraint loss L_{dk} in the dynamic keypoint box modulation module, we conduct a comparative study. Specifically, **w/o L** denotes the model variant without L_{dk}, while **w/ L** includes the additional loss. As illustrated in Fig. 4, the incorporation of L_{dk} leads to consistent improvements across all evaluation metrics, validating the effectiveness of the proposed constraint in guiding more accurate and adaptive keypoint box size modulation.

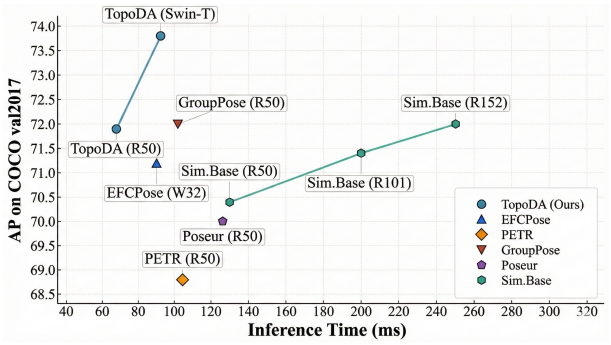

Figure 5: Comparison of inference time and performance trade-offs of different methods on the COCO val2017 set.

Comparison of Effectiveness

To further assess the efficiency of our method, we compare inference time and accuracy across recent approaches, as shown in Fig. 5. TopoDA achieves a strong trade-off, reaching 73.8 AP with a runtime of 93 ms. Compared to other single-stage methods, it reduces inference time by 33.3% while maintaining competitive accuracy. This efficiency stems from our streamlined architecture, which accelerates computation without sacrificing keypoint precision. These results highlight TopoDA's suitability for real-time applications such as human-computer interaction and autonomous perception, where both speed and accuracy are critical.

Conclusion

Effectively handling occlusion, scale variation, and complex human interactions remains a longstanding challenge in multi-person pose estimation. In this work, we introduce TopoDA, a topology-aware dynamic association framework that systematically addresses these issues through a coarse-to-fine refinement paradigm. By explicitly modeling person-to-keypoint relationships, dynamically modulating keypoint receptive fields, and incorporating cross-keypoint reasoning over flexible keypoint associations, TopoDA enables more flexible and robust pose inference under challenging conditions. Extensive experiments on COCO and CrowdPose benchmarks demonstrate that our method achieves state-of-the-art accuracy, particularly in densely populated scenes with severe occlusions. These results validate the effectiveness of our topology-aware design and highlight the practical advantages of integrating dynamic structural modeling into end-to-end pose estimation frameworks.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. The work was partially supported by the Major Project of Interdisciplinary Scientific Research Platform of Central China Normal University (No. CCNU25JCPT019), China Postdoctoral Science Foundation (No. 380962).

References

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chen, S.; Zhang, Y.; Huang, S.; Yi, R.; Fan, K.; Zhang, R.; Chen, P.; Wang, J.; Ding, S.; and Ma, L. 2024. Sdpose: Tokenized pose estimation via circulation-guide self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1082–1090.

Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T. S.; and Zhang, L. 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5386–5395.

Ding, Y.; Deng, W.; Zheng, Y.; Liu, P.; Wang, M.; Cheng, X.; Bao, J.; Chen, D.; and Zeng, M. 2022. Î2R-Net: Intra- and Inter-Human Relation Network for Multi-Person Pose Estimation. *arXiv preprint arXiv:2206.10892*.

Fang, H.-S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.-L.; and Lu, C. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7157–7173.

Farias, T. d. S.; and Maziero, J. 2023. Feature alignment as a generative process. *Frontiers in Artificial Intelligence*, 5: 1025148.

Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; and Wang, J. 2021. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14676–14686.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Huang, S.; Lu, Z.; Cheng, R.; and He, C. 2021. Fapn: Feature-aligned pyramid network for dense image prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 864–873.

Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; and Schieke, B. 2016. DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. In *European Conference on Computer Vision (ECCV)*.

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

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, B.; Tang, S.; and Li, W. 2024. LMFormer: Lightweight and multi-feature perspective via transformer for human pose estimation. *Neurocomputing*, 594: 127884.

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; and Lu, C. 2019. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10863–10872.

Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; and Tu, Z. 2021. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1944–1953.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Chen, Q.; Tan, Z.; Liu, J.-J.; Wang, J.; Su, X.; Li, X.; Yao, K.; Han, J.; Ding, E.; et al. 2023. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15029–15038.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8759–8768.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, P.; Jiang, T.; Li, Y.; Li, X.; Chen, K.; and Yang, W. 2024. 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*, 1491–1500.

Mao, W.; Tian, Z.; Wang, X.; and Shen, C. 2021. Fcpose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9034–9043.

Newell, A.; Huang, Z.; and Deng, J. 2017. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in Neural Information Processing Systems*, 30.

Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.; and Schiele, B. 2016. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, D.; Wei, X.; Li, L.; Ren, Y.; and Tan, W. 2022. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11069–11078.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790.

Toshev, A.; and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1653–1660.

Wang, D.; and Zhang, S. 2022. Contextual Instance Decoupling for Robust Multi-Person Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11060–11068.

Wang, D.; Zhang, S.; and Hua, G. 2021. Robust Pose Estimation in Crowded Scenes with Direct Pose-Level Inference. *Advances in Neural Information Processing Systems*, 34: 6278–6289.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 4724–4732.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 466–481.

Xue, N.; Wu, T.; Xia, G.-S.; and Zhang, L. 2022. Learning local-global contextual adaptation for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13065–13074.

Yang, J.; Zeng, A.; Liu, S.; Li, F.; Zhang, R.; and Zhang, L. 2023. Explicit box detection unifies end-to-end multi-person pose estimation. *arXiv preprint arXiv:2302.01593*.

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

Zhang, Y.; Fu, L.; Li, Y.; and Zhang, Y. 2021. HDFNet: Hierarchical dynamic fusion network for change detection in optical aerial images. *Remote Sensing*, 13(8): 1440.

Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; and Shah, M. 2023. Deep Learning-Based Human Pose Estimation: A Survey. *ACM Comput. Surv.*

Zhong, X.; Qin, J.; Guo, M.; Zuo, W.; and Lu, W. 2022. Offset-decoupled deformable convolution for efficient crowd counting. *Scientific Reports*, 12(1): 12229.