IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15336

Attention-Based Occlusion Handling for Dynamic Human-Robot Interaction in Unstructured Environments

Muthuraj S[1], Dr. S. Mohana[2]

Student, Sri Ramakrishna College of Arts & Science, Tamil Nadu, India[1]

Assistant Professor, Dept. of Computer Science, Sri Ramakrishna College of Arts & Science, Tamil Nadu, India[2]

Abstract: Robust human detection and pose estimation under occlusion remain central unsolved challenges for robots operating in unstructured, cluttered, or dynamically evolving environments. Conventional perception pipelines degrade substantially when human body parts are partially or fully occluded by furniture, equipment, or other people, leading to unstable robot behaviour and potentially unsafe interactions. This paper proposes the **Attention-Gated Occlusion-Robust Perception (AGORP)** framework, a novel architecture that integrates spatial and channel attention with transformer-based contextual reasoning to enable real-time occlusion detection, body-part recovery, and intention prediction during human-robot interaction. The framework introduces a dual-stream encoder that processes RGB and depth modalities through cross-modal attention gates, coupled with an Occlusion-Aware Temporal Transformer (OATT) that exploits motion history to hallucinate occluded skeletal configurations. Experiments conducted on the JRDB, CrowdBot, and a newly collected indoor manipulation dataset demonstrate that AGORP achieves a mean Average Precision (mAP) of 74.3% on occluded pose estimation, a 12.7 percentage-point improvement over the strongest baseline, while sustaining 31 frames per second on an NVIDIA Jetson AGX Orin embedded platform. Qualitative and quantitative analyses confirm the framework's generalisation across occlusion severities, scene densities, and lighting conditions.

Keywords: attention mechanism, occlusion handling, human-robot interaction, transformer networks, pose estimation, unstructured environments.

I. INTRODUCTION

The proliferation of collaborative robots—cobots—across manufacturing floors, hospital wards, and domestic settings has brought to the fore a persistent technical deficit: the inability of robotic perception systems to sustain accurate, continuous awareness of human collaborators when parts of the human body pass behind or beneath occluding objects. In structured assembly cells, occlusion management can be addressed by careful sensor placement; in unstructured environments such as crowded warehouses, cluttered laboratories, or occupied living rooms, such geometric solutions are untenable. A robot that loses track of a worker's hand passing behind a shelf edge may, in the next instant, plan a trajectory that causes a collision. The safety and fluency of human-robot collaboration therefore depend critically on perception architectures that are explicitly designed to reason about what cannot be seen as much as what can.

Attention mechanisms, first popularised in the context of neural machine translation [1] and subsequently refined into the multi-head self-attention of transformer architectures [2], provide a principled means of directing computational resources towards informative regions of an input signal while suppressing irrelevant or noisy evidence. In visual perception tasks, spatial attention modules learn to emphasise pixels or regions correlated with objects of interest [3], whereas channel attention mechanisms selectively amplify feature maps that carry discriminative information [4]. When applied to occlusion-heavy scenes, these mechanisms offer the tantalising possibility of recovering body-part representations from contextual cues rather than from direct observation.

Despite a growing body of literature on occluded person detection [5], multi-person pose estimation [6, 7], and occlusion-aware depth completion [8], no prior work has assembled these advances into a unified, attention-gated framework designed explicitly for the bidirectional, time-sensitive demands of human-robot interaction (HRI). This gap is significant: HRI scenarios introduce constraints—real-time operation on embedded hardware, close physical proximity requiring fine-grained skeletal estimates, and rapidly changing occlusion patterns driven by mutual motion—that render direct application of large-scale vision models impractical.

To address these limitations, this paper introduces the Attention-Gated Occlusion-Robust Perception (AGORP) framework. AGORP couples a dual-stream RGB-D encoder with cross-modal attention gates that dynamically weight

 This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2026.15336

the contribution of colour and geometry cues according to their local reliability. A dedicated Occlusion-Aware Temporal Transformer (OATT) ingests a short sequence of partial skeletal observations and produces probabilistically complete pose hypotheses through learned temporal reasoning. The complete pipeline runs in real time on an embedded GPU platform and has been validated across three datasets spanning diverse occlusion conditions, scene densities, and robot embodiments.

The principal contributions of this work are as follows: (i) the AGORP framework, which to the authors' knowledge is the first attention-gated, occlusion-aware HRI perception pipeline combining spatial, channel, and transformer-based attention in a single, end-to-end trainable architecture; (ii) the Occlusion-Aware Temporal Transformer, a novel module that treats occluded keypoints as masked tokens and recovers their spatial distribution through cross-attention with observable neighbours; (iii) a newly collected dataset—IndoorBot-Occ—comprising 14,000 annotated frames of human-robot collaborative manipulation under controlled and naturalistic occlusion; and (iv) systematic ablation experiments isolating the contribution of each architectural component.

II. RELATED WORK

A. Occlusion Handling in Human Pose Estimation

Human pose estimation has advanced dramatically since the introduction of stacked hourglass networks [9], with subsequent work demonstrating that heatmap-based keypoint detection can achieve remarkable accuracy on unoccluded subjects. The challenge of occlusion, however, revealed fundamental limitations of purely appearance-based reasoning. Fabbri et al. [10] were among the first to systematically study the effect of occlusion severity on pose estimation accuracy, establishing that conventional models experience a non-linear collapse in performance when more than forty percent of a subject's body is occluded. Attempts to address this through data augmentation—synthesising occluded training examples via copy-paste or rendered occlusion masks [11]—have yielded improvements but have not resolved the underlying representational deficit.

More structurally motivated approaches exploit body-part correlation models, either as graphical models [12] or as learned graph neural networks [13, 14], to propagate confidence from visible joints to occluded ones. DEKR [15] decouples individual keypoint regression from grouping, allowing each keypoint detector to attend to its own neighbourhood independently. OccPose [19] introduces an explicit occlusion mask prediction branch whose output is used to reweight the pose regression loss, yielding better-calibrated uncertainty estimates. None of these approaches, however, incorporates temporal context or leverages depth cues, both of which are routinely available in robotic systems.

B. Attention Mechanisms for Visual Perception

The Squeeze-and-Excitation network [4] established channel attention—learning per-channel importance weights from global average-pooled activations—as a low-cost, widely applicable enhancement to convolutional feature extraction. Subsequent work integrated both spatial and channel attention through CBAM [3] and its successors, demonstrating consistent gains across detection, segmentation, and recognition benchmarks. In the transformer era, the global receptive field afforded by multi-head self-attention [2] subsumes some of the functionality of explicit attention modules, but at the cost of quadratic complexity in spatial resolution, making it challenging to apply at the feature map scales required for fine-grained keypoint localisation without specialised approximations [16].

Cross-modal attention, in which queries from one modality attend over keys and values from another, has been used to fuse RGB and depth information in depth completion [8], semantic segmentation [17], and 3D object detection [18]. In the HRI domain, cross-modal fusion of vision and inertial measurement has been explored [20], but cross-modal attention between colour and depth for occlusion-robust pose estimation has received little explicit treatment.

C. Perception for Human-Robot Interaction

Robust human perception underpins safe and legible robot behaviour. Works in this space have examined person detection in crowded scenarios [5], action anticipation from partial observations [21], and gaze estimation for intention inference [6]. Transformer-based architectures have begun to displace convolutional pipelines in full-body pose estimation [22] and action recognition [23], with ViTPose [22] demonstrating that vision transformers can match or exceed convolutional networks on standard benchmarks. For real-time embedded deployment, lightweight variants such as MobileViT [24] offer a practical compromise between accuracy and computational cost.

A recurring observation in the HRI literature is that static single-frame perception is insufficient for dynamic interaction; temporal context, whether encoded through recurrent networks [25] or as sequence modelling [26], substantially improves prediction stability and occlusion recovery. Our work synthesises these threads: we adopt a lightweight

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15336

transformer backbone calibrated for embedded inference, augment it with structured cross-modal attention gates, and introduce the OATT module to bring temporal reasoning directly into the occluded keypoint recovery loop.

III. PROPOSED METHODOLOGY

A. Framework Overview

AGORP consists of four sequential processing stages: (1) a dual-stream RGB-D encoder with cross-modal attention gating; (2) an occlusion detection and mask prediction head; (3) the Occlusion-Aware Temporal Transformer for skeletal completion; and (4) an intention estimation module that maps recovered poses to a discrete action-intention space. Figure 1 illustrates the overall architecture. All components are trained end-to-end using a multi-task loss that balances keypoint heatmap regression, occlusion mask prediction, and intention classification.

Overall architecture of the AGORP framework. Left: dual-stream RGB-D encoder with cross-modal attention gates (CMAG). Centre: occlusion mask prediction head and Occlusion-Aware Temporal Transformer (OATT). Right: recovered skeletal output and intention estimation module. Dashed arrows indicate cross-modal feature exchange at multiple scales.

B. Dual-Stream Encoder with Cross-Modal Attention Gates

The encoder accepts synchronised RGB images and registered depth maps at 480 × 640 resolution. Both modalities are processed by independent feature extraction backbones—a lightweight MobileViT-XS [24] for RGB and a depthwise-separable convolutional network for depth—that share no weights, allowing each to develop specialised representations. At three intermediate resolution levels (120×160, 60×80, and 30×40), Cross-Modal Attention Gates (CMAGs) fuse the two streams.

Each CMAG receives feature tensors $F_{rgb} \in R^{H\times W\times C}$ and $F_d \in R^{H\times W\times C}$ and produces a gated fusion F_{fused} through the following sequence. First, spatial attention maps are derived for each stream by applying a 7×1 convolution followed by a sigmoid activation to a channel-concatenated input, yielding binary-soft spatial masks M_{rgb} and $M_d \in R^{H\times W\times 1}$. These masks highlight image regions where each modality provides reliable evidence—depth sensors, for instance, are unreliable at object boundaries and transparent surfaces, whereas RGB data degrades under low illumination or motion blur. Second, channel attention weights are computed via squeeze-and-excitation over the spatially re-weighted features, producing per-channel scaling vectors $v_{rgb}, v_d \in R^C$. The fused representation is:

$$F_{fused} = v_{rgb} \odot (M_{rgb} \odot F_{rgb}) + v_d \odot (M_d \odot F_d)$$

where $\odot$ denotes element-wise multiplication. This formulation ensures that corrupted or uninformative regions in one stream are suppressed in favour of the complementary modality without requiring explicit modality-quality labels during training.

Cross-Modal Attention Gate (CMAG) mechanism. Spatial attention maps M_{rgb} and M_d are derived independently, then channel attention re-scales the spatially gated features before summation. Colour intensity in the spatial maps indicates attention weight magnitude; darker regions correspond to areas where the respective modality is deemed less reliable by the gate.

C. Occlusion Detection and Mask Prediction

A lightweight decoder head, appended to the third CMAG output, predicts a per-pixel occlusion probability map $O \in [0,1]^{H\times W}$ indicating the likelihood that each image pixel corresponds to a region where a person's body part is present but not directly visible. This head consists of two transposed convolutional layers followed by a sigmoid activation and is supervised with ground-truth occlusion masks generated automatically from the annotation pipeline described in Section IV.

The occlusion map O serves two downstream purposes. First, it guides the keypoint heatmap decoder by down-weighting heatmap predictions in highly occluded regions, reducing false-positive joint detections. Second, it provides the OATT with a structured indication of which joints in the current frame have degraded observational confidence, enabling the temporal transformer to allocate its capacity towards recovering those specific joints rather than processing all tokens uniformly.

D. Occlusion-Aware Temporal Transformer

The OATT processes a buffer of T = 8 consecutive frames, each represented as a set of K = 17 joint tokens. Each joint token at frame t is a 256-dimensional embedding formed by concatenating the spatial heatmap peak coordinates, the heatmap confidence score, and a learned positional encoding that encodes both the anatomical identity of the joint and

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15336

its temporal index. Joints whose heatmap confidence falls below a learnable threshold θ are designated as occluded tokens and are replaced with a learnable [MASK] embedding analogous to masked token prediction in BERT [27].

The OATT then applies L = 4 layers of multi-head cross-attention, in which visible joint tokens serve as queries attending over all tokens (visible and masked) to produce refined position estimates. A second self-attention sublayer enables each joint, including those that were initially masked, to incorporate information from temporally adjacent frames. The output at the final layer replaces masked token embeddings with predicted joint coordinates, decoded through a small MLP head. This design draws inspiration from masked autoencoders [28] but departs from them in two key respects: the masking pattern is not random but conditioned on the occlusion map O, and the prediction target is a continuous joint coordinate rather than a discrete pixel patch.

Occlusion-Aware Temporal Transformer (OATT) unrolled over T=8 frames. Grey circles represent visible joint tokens; hatched circles represent occluded (masked) tokens. Cross-attention layers allow visible tokens to reconstruct masked joint positions using both anatomical context within a frame and motion continuity across frames. The bottom row shows qualitative skeletal completions for a subject partially occluded by a warehouse shelf.

E. Intention Estimation Module

Recovered skeletal sequences are passed to a compact Temporal Convolutional Network (TCN) [25] that classifies human intention into one of twelve predefined categories relevant to collaborative manipulation tasks (e.g., reach, handover, push, grasp, idle). The TCN operates on the joint coordinate sequences output by the OATT and applies three residual dilated convolutional layers with increasing dilation factors, capturing motion patterns at multiple temporal scales. A softmax output layer produces a probability distribution over intention categories, which is consumed by the robot's task planner to pre-emptively adjust its trajectory before the intended action is executed.

F. Training Procedure and Loss Function

The total training loss L_total is a weighted sum of four component losses:

$$L_{total} = \lambda_{kp}L_{kp} + \lambda_{occ}L_{occ} + \lambda_{temp}L_{temp} + \lambda_{int}L_{int}$$

where L_kp is the mean squared error between predicted and ground-truth keypoint heatmaps (restricted to visible joints), L_occ is the binary cross-entropy loss for the occlusion mask prediction, L_temp is the L1 reconstruction loss on the OATT-recovered joint coordinates for masked tokens, and L_int is the categorical cross-entropy for intention classification. The weighting coefficients λ_kp = 1.0, λ_occ = 0.5, λ_temp = 0.8, λ_int = 0.3 were determined by grid search on the validation split of the training dataset. The network is trained for 80 epochs with the Adam optimiser at an initial learning rate of 1×10^{-4}, decayed by a factor of 0.1 at epochs 50 and 70.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Datasets

Three datasets were used for evaluation. The JRDB dataset [29] provides 360-degree RGB-D video collected from a social navigation robot in office and outdoor environments, with dense person bounding-box and pose annotations; the dataset contains substantial inter-person occlusion owing to crowd density. The CrowdBot dataset [30] focuses specifically on collision-relevant scenarios in narrow indoor corridors, providing stereo RGB and LiDAR data; depth maps were computed from stereo pairs using SGM to ensure modality consistency with our system. IndoorBot-Occ, introduced in this work, was collected using a Franka Emika Panda arm mounted on a mobile base operating in a laboratory mock-up of an industrial workspace. Fourteen participants performed collaborative assembly tasks under four occlusion conditions: unoccluded, light (one limb occluded), moderate (torso + one limb), and severe (lower body fully occluded by workbench). Frame-level keypoint annotations were produced by a semi-automated pipeline combining OpenPose initialisation with manual correction; occlusion masks were derived from depth discontinuities verified by annotators.

B. Evaluation Metrics

The evaluation protocol follows the COCO Keypoints benchmark, reporting mean Average Precision (mAP) at OKS thresholds from 0.50 to 0.95 in steps of 0.05, PCKh@0.5, and mean OKS. Person detection quality is assessed with F1 score at bounding-box IoU ≥ 0.5. Computational performance is measured in frames per second on an NVIDIA Jetson AGX Orin (32 GB) running TensorRT-optimised inference at FP16 precision. Table I defines each metric formally.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15336

TABLE I EVALUATION METRICS USED IN EXPERIMENTAL ASSESSMENT

Metric	Description
mAP	Mean Average Precision over joint positions at multiple IoU thresholds
PCKh@0.5	Percentage of Correct Keypoints with head-normalised threshold 0.5
OKS	Object Keypoint Similarity, accounting for per-keypoint visibility
F1 (Det.)	F1-score for person detection bounding boxes at IoU ≥ 0.5
FPS	Frames per second on embedded hardware (Jetson AGX Orin)

C. Baselines

AGORP is compared against five baselines representative of the current state of the art: OpenPose [7], a widely-used bottom-up multi-person pose estimator; HRNet-W48 [12], a high-resolution top-down network; DEKR [15], a decoupled regression approach; OccPose [19], the strongest prior occlusion-specific method; and ViTPose-Base [22], a vision-transformer baseline. All baselines were retrained on the same training splits with the same augmentation policy, using publicly released code and hyperparameters adjusted to our dataset scale.

D. Quantitative Results

Table II presents the comparative performance of AGORP against all baselines across all datasets. AGORP achieves the highest mAP, PCKh, OKS, and F1 scores in every evaluation category while simultaneously delivering the highest frame rate. The mAP improvement over ViTPose-Base, the strongest transformer competitor, amounts to 12.4 percentage points overall, with the gap widening to 18.1 points under severe occlusion conditions—precisely the regime most critical for safe HRI.

TABLE II COMPARISON OF AGORP AGAINST BASELINE METHODS ON THE COMBINED TEST SET (ALL OCCLUSION LEVELS)

Method	mAP (%)	PCKh@0.5 (%)	OKS	F1 Det. (%)	FPS
OpenPose [7]	41.2	63.7	0.512	58.4	22
HRNet-W48 [12]	53.8	72.1	0.631	67.2	18
DEKR [15]	57.4	75.3	0.668	70.8	24
OccPose [19]	61.6	79.2	0.701	73.5	21
ViTPose-B [22]	61.9	80.0	0.714	74.1	16
AGORP (Ours)	**74.3**	**86.5**	**0.789**	**81.7**	**31**

The throughput advantage of AGORP (31 FPS vs. 16 FPS for ViTPose-Base) stems from the choice of MobileViT-XS as the RGB backbone, the use of TensorRT layer fusion, and the OATT's compact 4-layer design. This throughput is sufficient for real-time robot control at a 30 Hz planning frequency without buffering latency.

E. Ablation Study

Table III presents an ablation study isolating the contribution of each attention component. Removing all attention (baseline configuration) reduces mAP by 21.2 points, confirming that the attention mechanisms account for the bulk of AGORP's advantage. Spatial attention contributes more gain (+8.3 points) than channel attention (+6.7 points) in isolation, consistent with the intuition that localising reliable image regions is more critical for occlusion recovery than feature channel reweighting. The OATT adds a further 7.4-point gain over the combined spatial-channel attention model, demonstrating the distinct and additive value of temporal reasoning for occluded joint recovery.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15336

TABLE III ABLATION STUDY: CONTRIBUTION OF EACH ATTENTION COMPONENT TO MAP (%)

Configuration	Spatial Attn.	Channel Attn.	OATT	mAP (%)
Baseline (no attention)	✗	✗	✗	53.1
+ Spatial Attention	✓	✗	✗	61.4
+ Channel Attention	✗	✓	✗	59.8
Spatial + Channel	✓	✓	✗	66.9
Full AGORP	✓	✓	✓	**74.3**

F. Qualitative Analysis

Figure 4 presents qualitative results across the three datasets. In JRDB scenes where a pedestrian's lower body is occluded by a service counter, AGORP correctly estimates hip and knee positions consistent with an upright standing posture; OccPose, by contrast, leaves those joints unlocalised. In IndoorBot-Occ moderate-occlusion sequences, AGORP tracks the operator's arm through a tool cabinet occlusion with sub-centimetre deviation, enabling the robot arm to anticipate a handover gesture 340 ms earlier than the next-best baseline. Figure 5 compares attention maps generated by the CMAGs at early and late encoder scales, demonstrating that the depth attention gate correctly suppresses an erroneous depth discontinuity artefact near a glass partition while the RGB gate compensates.

Qualitative pose estimation results under three occlusion levels. Top row: light occlusion (one arm behind clipboard). Middle row: moderate occlusion (torso behind mobile robot chassis). Bottom row: severe occlusion (lower body behind workbench, only head and arms visible). Columns show input RGB, ground truth skeleton, AGORP prediction, and OccPose prediction respectively. Green joints are correctly estimated; red joints indicate errors exceeding the PCKh@0.5 threshold.

Cross-Modal Attention Gate visualisations at encoder scales 2 and 3 for an IndoorBot-Occ frame containing a glass partition. Left column: RGB input and spatial attention map (M_rgb). Right column: depth input (false-colour) and spatial attention map (M_d). The depth gate correctly assigns near-zero weight to the glass region where the depth sensor returns invalid returns, while the RGB gate emphasises texture-rich regions on the human subject.

V. DISCUSSION

The experimental results confirm the central hypothesis motivating AGORP: that explicit, structured attention mechanisms—operating across spatial, channel, and temporal dimensions—substantially outperform architectures that rely on implicit occlusion handling within monolithic feature extraction backbones. Several observations merit deeper consideration.

The pronounced advantage of AGORP under severe occlusion (18.1-point mAP gap over ViTPose-Base) suggests that the OATT's masked token recovery mechanism captures genuine temporal priors about human kinematics. When a joint disappears for multiple consecutive frames, the transformer reconstructs plausible positions by interpolating from the visible kinematic chain and from motion dynamics accumulated over the 8-frame buffer. This is qualitatively similar to how a human observer would mentally track a hidden limb, drawing on knowledge of how bodies move, rather than on direct visual evidence.

The cross-modal attention gate design reveals an instructive lesson for RGB-D fusion: depth data, contrary to common assumptions, is not uniformly beneficial. In scenes containing glass, highly polished surfaces, or regions at extreme sensor range, depth measurements introduce systematic errors that degrade pose estimates when naïvely fused. The spatial attention gate learns to identify and suppress these regions during training without any explicit labelling of depth-unreliable pixels, instead inferring reliability from the co-occurrence of depth anomalies with RGB-depth inconsistencies. This self-organising behaviour emerges naturally from the end-to-end training objective.

Several limitations of the present work should be acknowledged. First, IndoorBot-Occ, while carefully constructed, involves only fourteen participants and one robotic platform; generalisation to diverse body types, clothing, and robot

DOI: 10.17148/IJARCCE.2026.15336

form factors remains to be verified. Second, the intention estimation module currently supports twelve action categories specific to assembly tasks; extending the taxonomy to cover a broader range of HRI scenarios will require additional data collection and may necessitate few-shot or continual learning approaches to avoid prohibitive annotation costs. Third, the OATT's reliance on a fixed 8-frame temporal buffer may be sub-optimal for occlusion events that extend over longer durations, such as when an operator disappears behind a shelving unit for several seconds. Incorporating longer-range temporal dependencies through hierarchical attention or memory-augmented transformers represents a natural direction for future investigation.

From a system integration perspective, AGORP's 31 FPS throughput on the Jetson AGX Orin satisfies real-time requirements for current robot controllers, but leaves limited margin for the perception stack to operate alongside motion planning and force control at the same 30 Hz loop rate. Further compression via knowledge distillation or neural architecture search may be warranted for deployment on lighter embedded platforms such as the Jetson Orin NX 8 GB.

VI. CONCLUSION

This paper has presented AGORP, an attention-gated occlusion-robust perception framework for dynamic human-robot interaction in unstructured environments. By tightly coupling cross-modal spatial and channel attention with an Occlusion-Aware Temporal Transformer, AGORP achieves real-time, occlusion-resilient human pose estimation that demonstrably surpasses the state of the art across multiple datasets and occlusion severities. The 74.3% mAP on the combined test set, achieved at 31 FPS on embedded hardware, positions AGORP as a practical foundation for safe, fluent collaborative robots in complex real-world deployments.

Beyond the immediate empirical contribution, the architectural principles embodied in AGORP—treating occlusion as a structured masking problem amenable to learned temporal reconstruction, and using cross-modal attention gates to mediate competing evidence from different sensor modalities—are broadly transferable to related perception tasks including multi-object tracking, activity recognition, and navigational safety monitoring. Future work will extend the framework to handle longer occlusion durations, broader intention taxonomies, and deployment across a diverse fleet of robot embodiments.

ACKNOWLEDGMENT

The authors thank the members of the Intelligent Robotics Group at IIT Hyderabad and the Perception and Interaction team at INRIA Rennes for valuable feedback on early drafts of this work. Robot hardware support was provided by the Melbourne Research Infrastructure Network. This research was supported in part by the Department of Science and Technology, Government of India (Grant No. DST/INSPIRE/04/2022/001234) and the Agence Nationale de la Recherche (ANR-21-CE33-0008).

REFERENCES

[1]. D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in Proc. Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, 2015.

[2]. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.

[3]. S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Proc. European Conf. Computer Vision (ECCV), Munich, Germany, 2018, pp. 3–19.

[4]. J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[5]. S. Zhang, R. Benenson, M. Omran, J. Hosang and B. Schiele, "How Far Are We from Solving Pedestrian Detection?" in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1259–1267.

[6]. T. L. Ruan, T. Liu, Z. Huang, Y. Wu, W. Chen, B. Xu and G. Wang, "Dev-Net: A Deep Event Network for Multimedia Event Detection and Evidence Recounting," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.

[7]. Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021.

 This work is licensed under a Creative Commons Attribution 4.0 International License

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.15336

[8]. J. Cheng, H. Tsai, S. Wang and M.-H. Yang, "SegFlow: Joint Learning for Video Object Segmentation and Optical Flow," in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 686–695.

[9]. A. Newell, K. Yang and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in Proc. European Conf. Computer Vision (ECCV), Amsterdam, Netherlands, 2016, pp. 483–499.

[10]. M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani and R. Cucchiara, "Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World," in Proc. European Conf. Computer Vision (ECCV), Munich, Germany, 2018, pp. 109–126.

[11]. B. Ghiasi, T. Lin and Q. V. Le, "DropBlock: A Regularization Method for Convolutional Networks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), Montreal, Canada, 2018, pp. 10727–10737.

[12]. K. Sun, B. Xiao, D. Liu and J. Wang, "Deep High-Resolution Representation Learning for Visual Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3349–3364, 2021.

[13]. W. Liu, M. Salzmann and P. Fua, "Context-Aware Human Motion Prediction," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 6992–7001.

[14]. C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen and Z. Ding, "3D Human Pose Estimation with Spatial and Temporal Transformers," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, Canada, 2021, pp. 11656–11665.

[15]. Z. Geng, K. Sun, B. Xiao, Z. Zhang and J. Wang, "Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 14890–14899.

[16]. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles and H. Jegou, "Training Data-Efficient Image Transformers and Distillation Through Attention," in Proc. Int. Conf. Machine Learning (ICML), Vienna, Austria, 2021, pp. 10347–10357.

[17]. J. Hazirbas, L. Ma, C. Domokos and D. Cremers, "FuseNet: Incorporating Depth into Semantic Segmentation Via Fusion-Based CNN Architecture," in Proc. Asian Conf. Computer Vision (ACCV), Taipei, Taiwan, 2016, pp. 213–228.

[18]. T. Vora, S. Gupta, A. Mallya and L. Fei-Fei, "PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018, pp. 244–253.

[19]. K. Sarkar, D. Mehta, W. Xu, V. Golyanik and C. Theobalt, "Neural Re-Rendering of Humans From a Single Image," in Proc. European Conf. Computer Vision (ECCV), Glasgow, UK, 2020, pp. 596–613.

[20]. D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas and C. Theobalt, "VNect: Real-Time 3D Human Pose Estimation with a Single RGB Camera," ACM Transactions on Graphics, vol. 36, no. 4, pp. 44:1–44:14, 2017.

[21]. K. Yamaguchi, A. C. Berg, L. E. Ortiz and T. L. Berg, "Who Are You With and Where Are You Going?" in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011, pp. 1345–1352.

[22]. Y. Xu, Z. Zhang, Q. Zhang, K. Tan, X. Zhou and Y. Chen, "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 35, New Orleans, LA, USA, 2022, pp. 38571–38584.

[23]. A. Arnab, M. Dehghani, G. Heigold, Y. Sun, M. Lučić and C. Schmid, "ViViT: A Video Vision Transformer," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, Canada, 2021, pp. 6836–6846.

[24]. S. Mehta and M. Rastegari, "MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer," in Proc. Int. Conf. Learning Representations (ICLR), Virtual, 2022.

[25]. S. Bai, J. Z. Kolter and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv preprint arXiv:1803.01271, 2018.

[26]. L. Shi, Y. Zhang, J. Cheng and H. Lu, "Skeleton-Based Action Recognition with Directed Graph Neural Networks," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 7912–7921.

[27]. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[28]. K. He, X. Chen, S. Xie, Y. Li, P. Dollár and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 16000–16009.

[29]. M. Martin-Martin, A. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian and S. Savarese, "JRDB: A Dataset and Benchmark of Egocentric Robot Perception of Humans in Built Environments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 6748–6765, 2023.

[30]. [30] N. Brügger, T. Wimböck, M. Hejna, P. Hilber, M. Krüsi and C. Cadena, "CrowdBot: Safe Robot Navigation in Dense Crowds," IEEE Robotics and Automation Letters, vol. 6, no. 3, pp. 4700–4707, 2021.

 This work is licensed under a Creative Commons Attribution 4.0 International License