HOIMamba: Efficient Mamba-based Disentangled Progressive Learning for HOI Detection

Yongchao Xu [1], Jiawei Liu [1] *, Sen Tao [2], Qiang Zhang [1], Zheng-Jun Zha [1]

[1]University of Science and Technology of China, Hefei, China
[2] University of Chinese Academy of Sciences, Beijing, China
{yongchaoxu, zq_126 }@mail.ustc.edu.cn, taosen23@mails.ucas.ac.cn, {jwliu6, zhazj}@ustc.edu.cn

Abstract

Human-Object Interaction (HOI) detection aims to detect the spatial positions of human-object pairs and recognize their interactions. Existing single-branch, two-branch, and three-branch methods are challenging to make an appropriate trade-off on efficiency, multi-task decoupling, and collaborative learning, while they fail to identify rare and complex interaction categories effectively as well. In this work, we propose a novel Efficient Mamba-based Disentangled Progressive Learning (HOIMamba) for HOI Detection to absorb the advantages of the existing three approaches and adaptively aggregate multi-level interaction semantics guided by cross-task bidirectional information contexts. Specifically, HOIMamba builds an efficient and effective decoder through cascaded Low-Rank Adaptations (LoRAs), with high efficiency, thorough decoupling of tasks, and good multi-task collaborative learning. Furthermore, to alleviate the recognition problem of interactions in difficult HOI samples, a novel Mamba-based comprehensive progressive learning strategy with Cross-enhance Mamba (CEM) blocks and Detection Context Propagation (DCP) blocks is designed to gradually excavate interaction-related discriminative cues from four levels. CEM blocks automatically aggregate context to generate diverse task-shared semantics and simultaneously realize the cross-task interaction between human and object branches, guiding the interaction branch to extract more expressive HOI representation. DCP blocks further transfer the comprehensive interaction context to human and object branches to achieve rich and effective information exchange, facilitating the model to discover more HOI instances. Extensive experimental results on two standard benchmarks demonstrate the effectiveness of our HOIMamba.

Introduction

Human-Object Interaction (HOI) detection aims to detect human-object pairs and identify their interactions simultaneously. An increasing interest has been attracted to HOI detection due to its important role in a wide range of computer vision applications, such as 3D interaction understanding (Yang et al. 2023, 2024b) and image retrieval (Liu et al. 2019; Hu et al. 2024). The goal of the HOI detector is to localize all HOI samples within an image, representing the results in the form of triplets ⟨*human, verbs, object*⟩.

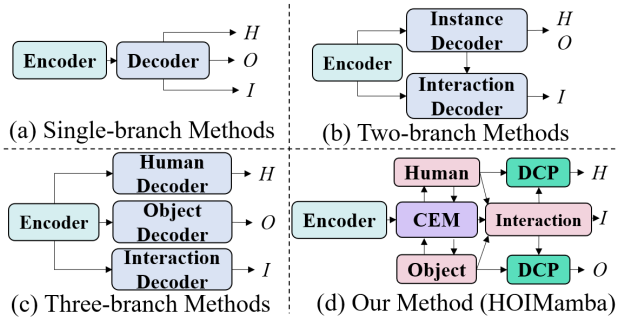

Figure 1: Comparison of HOIMamba and existing Transformer-based HOI methods, HOIMamba well absorbs the advantages of all paradigms. Moreover, the Mamba-based comprehensive progressive learning strategy is introduced to solve the gradual process of interaction learning that the existing methods lack, which greatly improves the performance of the HOI detector.

Previous HOI detection methods can generally be divided into two paradigms: CNN-based models (Wan et al. 2019; Gao et al. 2020) and Transformer-based models (Kim et al. 2021; Ning et al. 2023). Limited by the local receptive field of the convolution kernel, CNN-based methods are difficult to recognize complex HOIs. Thanks to the advantages of transformer in long-distance modeling, Transformer-based architecture has become the mainstream of HOI task development, and its performance has been greatly improved compared with CNN-based methods.

Existing Transformer-based methods can be classified into ***single-branch*** (Tamura, Ohashi, and Yoshinaga 2021), ***two-branch*** (Ning et al. 2023), and ***three-branch*** methods (Fang et al. 2023) according to the number of their decoder branches. As shown in Figure 1 (a), the ***single-branch methods*** use one decoder to be responsible for all sub-tasks (human detection, object detection, interaction prediction) simultaneously. Although the structure is simple, it is difficult for a single model to achieve a good trade-off in multi-task learning (Zhang et al. 2021). The ***two-branch methods*** decompose the HOI task into two sub-tasks: instance detection and interaction classification, and adopt two decoders to deal with these tasks respectively. As shown in Figure 1 (b), the

*Corresponding author.

two-branch methods concatenate the two decoders so that the human-object detection output is used as the initial query embedding of the interaction decoder, significantly enhances the collaborative learning of the model. Also, the two-branch methods achieve decoupling of human-object pairs detection and interaction recognition, but there are still coupling relationships between humans and objects, which constrains the possibility of discovering more HOI instances. To solve the problem of incomplete disentangling of two-branch, the ***three-branch methods*** completely decouple the three branches of human, object and interaction. As shown in Figure 1 (c), the three-branch methods explicitly utilize different parameters for human detection, object detection, and interaction classification, respectively, which further facilitates multi-task decoupling and collaborative learning. However, the interaction branch lacks valuable prior knowledge, making the model converge slowly. In addition, introducing the additional branch causes a large number of parameters, which brings a huge training cost and restricts the further development of HOI detection.

Through the above analysis, we can find the existing methods have their own advantages and disadvantages. Therefore, the intuitive idea is to take the essence and discard the dross. To achieve this, we propose HOIMamba, a novel HOI detection framework (Figure 1 (d)), to effectively combine the advantages of these paradigms via a Mamba-based decoder. Specifically, (1) we split the task into human detection, object detection, and interaction classification to **absorb the advantages of thoroughly decoupling within the three-branch methods**. (2) Also, we design each branch as a cascaded LoRAs architecture with fewer parameters to **maintain the simplicity of the single-branch methods**. (3) In addition, we dynamically weight the human and object features as part of the initial information of the interaction branch to provide a good prior for multi-task collaborative learning, **which takes the pros of the two-branch methods**. Moreover, in contrast to the three-branch methods that only explicitly separate the model parameters for three tasks respectively, HOIMamba sets personalized branches to each task to explicitly isolate different parameters, and trains the specific router for each task to dynamically assign weight combinations, achieving implicit gradient separation. The conflict between tasks is effectively alleviated **from both explicit and implicit perspectives.**

Furthermore, previous methods ignore that **identifying rare and complex interactions needs to capture more advanced visual semantics**. Therefore, to further improve the HOI detection performance, we introduce a novel Mamba-based comprehensive progressive learning strategy guided by cross-task bidirectional information contexts to gradually extract interaction features from low, middle, high, and comprehensive levels, and promote cross-task information exchange between three branches. The core compositions of the comprehensive progressive learning strategy are Cross-enhance Mamba (CEM) blocks and Detection Context Propagation (DCP) blocks based on Mamba (Gu and Dao 2023). CEM blocks adaptively aggregate multi-view task-shared semantics to ensure sufficient acquisition of comprehensive interaction features and simultaneously realize the cross-task interaction between human and object branches, while DCP blocks further transfer the comprehensive interaction context to human and object branches to achieve rich and effective information exchange. In general, the comprehensive progressive learning strategy first adopts the CEM blocks to guide the joint enhancement of the three branches. Then, we utilize the comprehensive interaction semantics within the interaction branch to reversely guide the feature learning of the human and object branches via the DCP blocks. The two kinds of blocks combined with cascade LoRAs to construct a comprehensive three-branch network to generate more expressive HOI representations from four levels.

Our contribution can be summarized as follows: (1) We introduce HOIMamba, a novel HOI detection framework, which includes a Mamaba-based decoder to effectively combine the advantages of single-branch, two-branch, and three-branch methods. (2) We design a novel comprehensive progressive learning strategy to facilitate the recognition of rare and complex interaction categories, whose core compositions are Cross-enhance Mamba (CEM) blocks and Detection Context Propagation (DCP) blocks for more expressive HOI representation. (3) Our method outperforms previous state-of-the-art methods by a large margin on two public benchmarks HICO-DET and V-COCO, and effectively reduces the computational complexity of the model.

Related Works

Transformer-based HOI Methods

Transformer-based methods rely on the self-attention and cross-attention mechanism (Carion et al. 2020) to effectively aggregate context information for HOI detection. These works can be divided into three categories according to the number of decoder branches: single-branch (Tamura, Ohashi, and Yoshinaga 2021), two-branch (Ning et al. 2023), and three-branch (Kim, Jung, and Cho 2023). The single-branch method uses a single transformer decoder to achieve object detection and relationship classification of human-object pairs simultaneously. The two-branch method decouples object detection and interaction recognition based on the single-branch method. Kim *et al*. (Kim et al. 2021) proposed HOTR based on a variation network, which consists of CNN, shared encoder, and parallel decoders. In the three-branch model, the relationship between human, object, and interaction is completely decoupled, and each decoder is responsible for different subtasks. Kim *et al*. (Kim, Jung, and Cho 2023) set up three branches in detection to drive relation inference by propagating valid context information.

State Space Models

The Structured State Space Sequence (S4) model (Gu, Goel, and Ré 2021) is first proposed to model long-range dependency based on the linear time-invariant system. The S4 model achieves excellent results on the long-range arena benchmarks for images, text, and speech, and its potential as a generalized sequence model deserves further exploration. Gu *et al*. (Gu and Dao 2023) proposed Mamba with selective SSMs based on the S4 model to implement dynamic weights

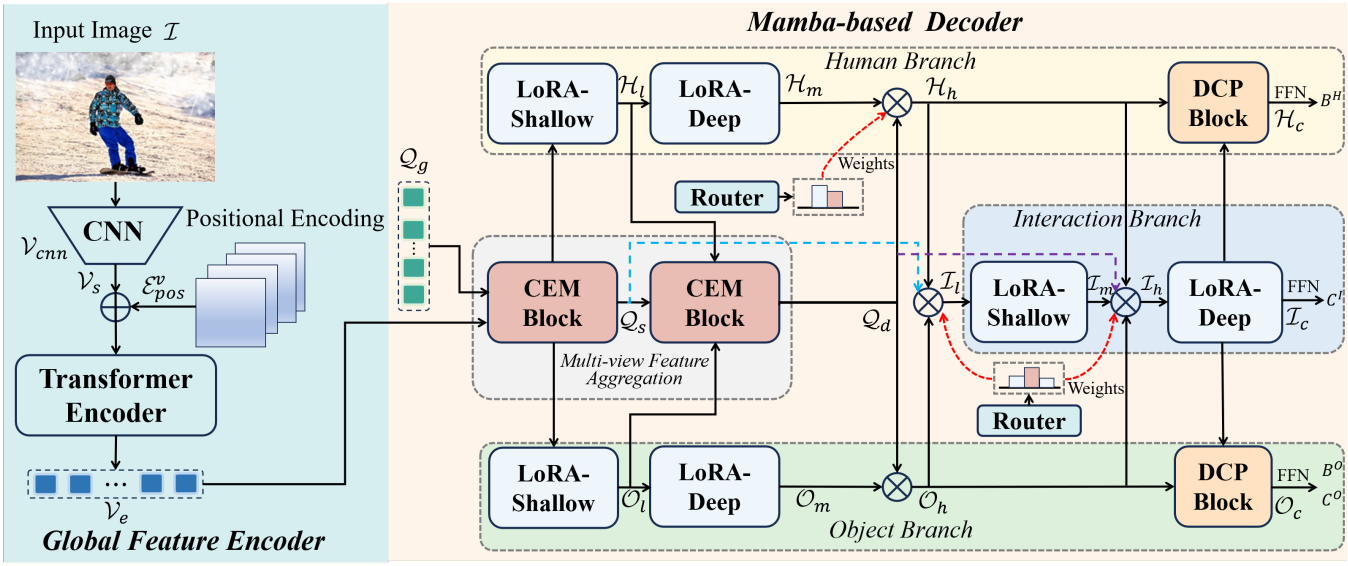

Figure 2: The overall pipeline of HOIMamba. It consists of one Global Feature Encoder and one Mamba-based Decoder.

in the field of NLP, which implies more flexibility and efficiency when dealing with diverse input. Since then, Mamba has been widely transferred to many computer vision tasks, and has achieved remarkable results in object classification (Liu et al. 2024; Zhao et al. 2024), image restoration (Guo et al. 2024) and other tasks. Guo *et al*. (Guo et al. 2024) introduced a Residual State-Space Block, and added a prior for image restoration tasks.

Low-Rank Adaptation

Low-Rank Adaptation (LoRA) (Hu et al. 2021) is a parameter-efficient fine-tuning method for pre-trained large models. It transfers large models to downstream tasks via training only a small number of additional parameters and keeping most of the pre-trained parameters frozen, reducing the training cost, and becoming a prevalent method of transfer learning. As one of the representatives, LoRAMoE (Dou et al. 2023) achieves an effective combination of LoRA and MoE methods, but its backbone and LoRA are still separated. In this work, we take a different perspective and aim to simplify the network with LoRA, facilitating the learning of task-specific information by incorporating both LoRA and the backbone network into the training phase.

Method

The proposed goal of HOIMamba is to mine the benefits of the existing paradigms and effectively identify rare and complex interaction categories. In this section, we first introduce the preliminary concepts related to Mamba, including the state space model and the discretization process of parameters. We then present the general framework of the HOIMamba model. Finally, we provide a comprehensive discussion of the entire HOIMamba architecture.

Preliminaries

State Space Model. Inspired by the continuous system modeling method in the classical control field, the state space model (Gu, Goel, and Ré 2021) is mainly built on the linear time-invariant system. By describing the hidden state and predicting the subsequent state of the input, the state space model is finally mapped to the output as the model representation. Mathematically, we represent the input one-dimensional function or sequence as $x(t)$ and convert it into the output $y(t)$ through the intermediate hidden state $h(t)$, which can be described by the linear Ordinary Differential Equation (ODE), as shown in Equation 1, where $\mathcal{A} \in \mathbb{R}^{N\times N}$, $\mathcal{B} \in \mathbb{R}^{N\times 1}$, $\mathcal{C} \in \mathbb{R}^{1\times N}$, and $\mathcal{D} \in \mathbb{R}^1$ are the parameters of the model.

$$h'(t) = \mathcal{A}h(t) + \mathcal{B}x(t), \tag{1}$$

$$y(t) = \mathcal{C}h(t) + \mathcal{D}x(t). \tag{2}$$

Parameter Discretization. State Space Model (SSM) cannot adapt to the requirements of deep learning scenarios due to its continuous system limitation. S4 and Mamba introduce discretization processing based on the state space model and realize the discretization of parameters $\mathcal{A}$ and $\mathcal{B}$ through time scaling parameter Δ, which can be discretized by zero-order preservation rules as follows:

$$\bar{\mathcal{A}} = \exp(\Delta\mathcal{A}), \tag{3}$$

$$\bar{\mathcal{B}} = (\Delta\mathcal{A})^{-1}\left(\exp(\Delta\mathcal{A}) - I\right) \cdot \Delta\mathcal{B}. \tag{4}$$

After the discretization of parameters $\mathcal{A}$ and $\mathcal{B}$ is implemented, Equations 1 and 2 can be converted into the discretization equations as follows:

$$h_k = \bar{\mathcal{A}}h_{k-1} + \bar{\mathcal{B}}x_k, \tag{5}$$

$$y_k = \bar{\mathcal{C}}h_k + \bar{\mathcal{D}}x_k. \tag{6}$$

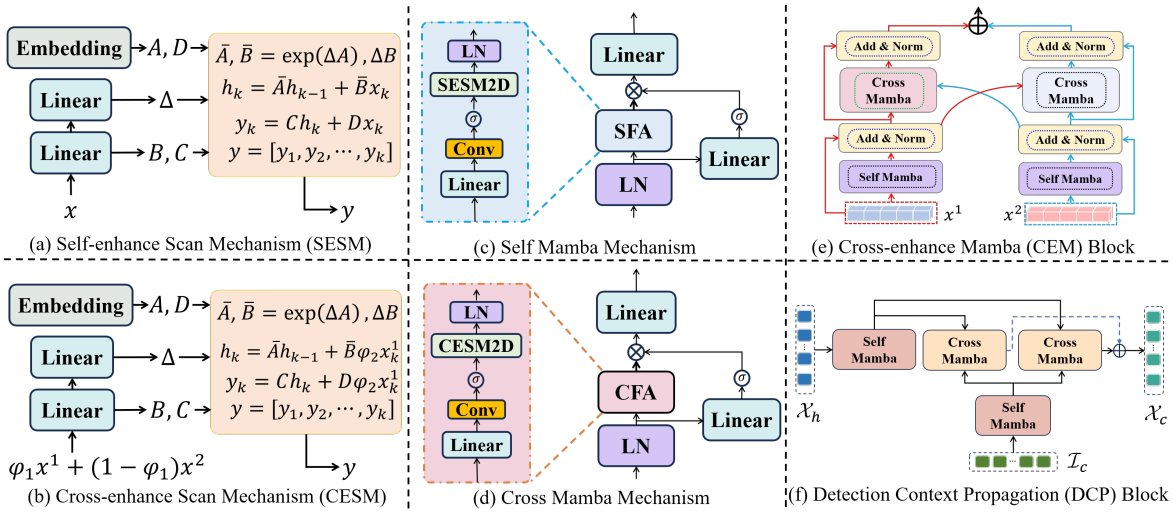

Figure 3: Illustration of the Cross-enhance Mamba (CEM) Block and Detection Context Propagation (DCP) Block.

The parallel calculation of the above process through global convolution can be expressed as follows:

$$\bar{K} = (\mathcal{CB}, \mathcal{CAB}, \ldots, \mathcal{CA}^{M-1}\mathcal{B}), \tag{7}$$

$$y = x * \bar{K}, \tag{8}$$

where $\bar{K}$ denotes the structured convolution kernel and M is the length of the input. The sequence output is generated by the parallel convolution operation, which effectively improves the calculation speed and scalability of the model.

Global Feature Encoder

Following the existing Transformer-based HOI methods, we adopt a CNN-transformer combined global feature encoder. For an input image $\mathcal{I} \in \mathbb{R}^{H\times W\times C}$, we first use CNN to obtain the visual feature $\mathcal{V}_{cnn} \in \mathbb{R}^{h\times w\times c}$, and reduce its channel dimension from c to C_e by a 1×1 convolution. Then we perform patch embedding by flattening operator to obtain the feature $\mathcal{V}_s \in \mathbb{R}^{(h\times w)\times C_e}$. The $\mathcal{V}_s$ is combined with the positional encoding $\mathcal{E}^v_{pos} \in \mathbb{R}^{(h\times w)\times C_e}$ and feed into the Transformer encoder to extract the global visual representation $\mathcal{V}_e \in \mathbb{R}^{(h\times w)\times C_e}$ for the following tasks.

Multi-view Feature Aggregation

To exploit global contextual features to enhance HOI pair representation, we propose a novel Cross-enhance Mamba (CEM) block to adaptively aggregate shallow and deep task-shared semantics from the global visual representation.

As shown in Figure 3, the original Mamba scan mechanism (Figure 3 (a)) is essentially a self-enhance scan mechanism ($SESM$), whose parameters Δ, B and C are obtained from the input x after linear transformation. Inspired by Vision Mamba (Zhu et al. 2024), $SESM$ can be combined with the bidirectional scan mechanism ($SESM2D$) to form the $SelfMamba$ mechanism (Figure 3 (c)). However, it can not realize adaptive aggregation for the HOI context well. Considering this, we introduce a novel scan mechanism called cross-enhance scan mechanism ($CESM$, as shown in Figure 3 (b)), which is mainly inspired by the idea that convolution kernels with specific parameters (Equation 7) can extract a comprehensive visual representation for specific needs (Equation 8). The $CESM$ can accept two inputs x^1 and x^2, in which case the parameters Δ, B and C will be obtained from the new inputs by a linear transformation.

$$linear(\varphi_1 x^1 + (1-\varphi_1)x^2) \longrightarrow \Delta, B, C, \tag{9}$$

where φ_1 is the hyperparameter that controls the mixing of the two inputs. Accordingly, Equations 5 and 6 will become as follows:

$$h_k = \bar{\mathcal{A}} h_{k-1} + \bar{\mathcal{B}}\varphi_2 {x_k}^1, \tag{10}$$

$$y_k = \bar{\mathcal{C}} h_k + \bar{\mathcal{D}}\varphi_2 {x_k}^1, \tag{11}$$

where φ_2 controls the input weights.

Similar to the $SelfMamba$, we combine $CESM$ with the bidirectional scan mechanism ($CESM2D$) to obtain the $CrossMamba$ mechanism (Figure 3 (d)). Then we design a two-tower structure to aggregate features, namely the Cross-enhance Mamba (CEM) block (Figure 3 (e)), which we denote as $CEM(x^1,x^2)$. Firstly, the two inputs x^1 and x^2 are enhanced by $SelfMamba$ respectively, and then they are fed into $CrossMamba$ to achieve multi-view feature aggregation. Finally, we integrate these to obtain the final output.

In the multi-view feature aggregation, we adopt two cascaded CEM blocks to obtain shallow and deep task-shared semantics respectively. Specifically, in the first CEM block, we feed the global visual representation to generate shallow task-shared semantic information $\mathcal{Q}_s \in \mathbb{R}^{N_q\times C_q}$:

$$\mathcal{Q}_s = CEM(\mathcal{Q}_g, \mathcal{V}_e), \tag{12}$$

where $\mathcal{Q}_g \in \mathbb{R}^{N_q\times C_q}$ denote the set of initial queries.

In the second CEM block, we get deep task-shared semantic information $\mathcal{Q}_d \in \mathbb{R}^{N_q\times C_q}$ by introducing the low-level human features $\mathcal{H}_l \in \mathbb{R}^{N_q\times C_q}$ and low-level object features $\mathcal{O}_l \in \mathbb{R}^{N_q\times C_q}$ within *Human and Object Branches Section*, which exchange context within the second CEM block, and further enhancing the comprehensive visual representation:

$$\mathcal{Q}_d = CEM(\mathcal{Q}_s, (\mathcal{H}_l + \mathcal{O}_l)/2). \tag{13}$$

Human and Object Branches

We design a four-stage feature extraction process for human and object branches to discover more HOI instances. In the first two stages, we use two cascaded Low-Rank Adaptations (LoRAs) to learn the low-level and middle-level human and object detection features respectively:

$$\Phi_{\theta\tau}(\mathcal{P}) = \mathcal{P}\mathcal{W}^{\tau}_{down}\mathcal{W}^{\tau}_{up}, \tau \in \{\mathcal{LS}, \mathcal{LD}\}, \quad (14)$$

where $\tau \in \{\mathcal{LS}, \mathcal{LD}\}$ represents LoRA-Shallow and LoRA-Deep, respectively, $\mathcal{W}^{\tau}_{down} \in \mathbb{R}^{C_q \times r}$ and $\mathcal{W}^{\tau}_{up} \in \mathbb{R}^{r \times C_q}$ are tunable parameters within the LoRA.

Specifically, through two cascaded LoRAs, low-level detection features $\mathcal{X}_l = \Phi_{\theta_{\mathcal{LS}}}(\mathcal{Q}_s) \in \mathbb{R}^{N_q \times C_q}$ and middle-level detection features $\mathcal{X}_m = \Phi_{\theta_{\mathcal{LD}}}(\mathcal{X}_l) \in \mathbb{R}^{N_q \times C_q}$ can be obtained in the first two stages, where $\mathcal{X} \in \{\mathcal{H}, \mathcal{O}\}$ indicates human and object respectively.

Next, in the third stage, we introduce a router $\Phi_{\theta\mathcal{XR}}$ for human and object branches respectively to dynamically weight middle-level detection features and deep task-shared semantics to obtain high-level detection features $\mathcal{X}_h \in \mathbb{R}^{N_q \times C_q}$ for human and object branches:

$$\mathcal{X}_h = \Phi_{\theta\mathcal{XR}}(\mathcal{X}_m, \mathcal{Q}_d) = \mathcal{W}^h_1\mathcal{X}_m + (1 - \mathcal{W}^h_1)\mathcal{Q}_d, \quad (15)$$

where $\mathcal{W}^h_1$ refers to the weight value within the router.

Finally, the fourth stage of human and object feature representation will be implemented in the *Detection Context Propagation Block Section*.

Interaction Branch

Improving the HOI performance on rare and complex interactions is crucial for practical applications. Therefore, inspired by the fact that the human brain needs a gradual process to learn complex things, we design a comprehensive progressive learning strategy for interaction classification to obtain more advanced high-level visual understanding.

Specifically, we first dynamically weight the shallow task-shared semantics and the high-level detection features in the first stage via a router $\Phi_{\theta\mathcal{IS}}$ to generate the low-level interaction features $\mathcal{I}_l = \Phi_{\theta\mathcal{IS}}(\mathcal{Q}_s, \mathcal{H}_h, \mathcal{O}_h) \in \mathbb{R}^{N_q \times C_q}$:

$$\Phi_{\theta\mathcal{IS}}(s_1, s_2, s_3) = \sum_{j=1}^{3} \mathcal{W}^{is}_j s_i, \sum_{j=1}^{3} \mathcal{W}^{is}_j = 1. \quad (16)$$

Next, the middle-level interaction features $\mathcal{I}_m = \Phi_{\theta_{\mathcal{LS}}}(\mathcal{I}_l) \in \mathbb{R}^{N_q \times C_q}$ are obtained by a LoRA-Shallow. Different from the human and object branches, we introduce a comprehensive progressive learning strategy and use an additional router $\Phi_{\theta\mathcal{ID}}$ for multi-view aggregation of semantics to obtain high-level interaction features $\mathcal{I}_h = \Phi_{\theta\mathcal{ID}}(\mathcal{Q}_d, \mathcal{H}_h, \mathcal{O}_h, \mathcal{I}_m) \in \mathbb{R}^{N_q \times C_q}$:

$$\Phi_{\theta\mathcal{ID}}(d_1, d_2, d_3, d_4) = \sum_{k=1}^{4} \mathcal{W}^{id}_k d_k, \sum_{k=1}^{4} \mathcal{W}^{id}_k = 1. \quad (17)$$

Furthermore, the comprehensive-level interaction features $\mathcal{I}_c = \Phi_{\theta_{\mathcal{LD}}}(\mathcal{I}_h) \in \mathbb{R}^{N_q \times C_q}$ is obtained by a LoRA-Deep.

Detection Context Propagation Block

In order to better realize information exchange, as shown in Figure 3 (f), we design a Detection Context Propagation (DCP) block to transfer the interaction content to the human and object branches. The DCP block consists of two $SelfMamba$ and $CrossMamba$, which further adaptively aggregates detection context from interaction information to increase detection performance, getting comprehensive-level detection features $\mathcal{X}_c = DCP(\mathcal{X}_h, \mathcal{I}_c) \in \mathbb{R}^{N_q \times C_q}$.

Finally, we embed the output of the three branches into the $\mathcal{FFN}s$ to generate a set of HOI predictions:

$$[\mathcal{B}^H; \mathcal{B}^O, \mathcal{C}^O; \mathcal{C}^I] = \mathcal{FFN}s([\mathcal{H}_c; \mathcal{O}_c; \mathcal{I}_c]), \quad (18)$$

where $\mathcal{B}^H \in \mathbb{R}^{N_q \times 4}$, $\mathcal{B}^O \in \mathbb{R}^{N_q \times 4}$ denote the bounding boxs of human and object respectively, $\mathcal{C}^O \in \mathbb{R}^{N_q \times N_o}$, $\mathcal{C}^I \in \mathbb{R}^{N_q \times N_i}$ indicate the categories of object and interaction respectively.

Model Training Objectives and Inference

Training. We follow the Transformer-based methods, treating the HOI detection task as a set prediction problem and using the Hungarian algorithm for bipartite graph matching. The loss function consists of boundary box regression loss L1 loss (Ren et al. 2015) and GIOU loss (Rezatofighi et al. 2019), cross-entropy loss of object classification, and focal loss (Lin et al. 2017) of interaction classification. The total loss function can be expressed as:

$$\mathcal{L} = \lambda_1 \sum_{i \in (h,o)} \mathcal{L}^i_{L1} + \lambda_2 \sum_{j \in (h,o)} \mathcal{L}^j_{GIoU} + \lambda_3 \mathcal{L}_{oc} + \lambda_4 \mathcal{L}_{ic}, \quad (19)$$

where λ_1, λ_2, λ_3 , and λ_4 are the hyperparameters to adjust the weight of each loss respectively.

Inference. At inference time, we add the object score from the object branch and the interaction score from the interaction branch to determine the confidence score for the prediction. Finally, the confidence scores are sorted, and the n with the highest score is taken as the final prediction result.

Experiments

Experimental Settings

Datasets and Evaluation Metrics. We evaluate the proposed model on two public benchmarks, HICO-DET (Chao et al. 2018) and V-COCO (Gupta and Malik 2015), and use the Mean Average Precision (mAP) metric on both datasets. A detailed description of the datasets and evaluation metrics can be found in the ***Supplementary Materials.***

Implementation Details

For our progressive learning, the hyperparameters for $CESM$ φ_1 and φ_2 are set to 0.50 and 0.75, respectively. During training, following MUREN (Kim, Jung, and Cho 2023), we set the number of queries N_q to 64, the number of channels C_e and C_q to 256, and the weight of the loss λ_1, λ_2, λ_3, and λ_4 are set to 3, 1, 1.25, and 1, respectively. The global feature encoder is initialized with DETR (Carion et al. 2020) parameters pre-trained on MS-COCO (Lin et al. 2014). We optimize the network by AdamW (Loshchilov

Method	Backbone	HICO-DET						V-COCO	
		Default			Known Object			Default	
		Full	Rare	Non-rare	Full	Rare	Non-rare	AP^{S1}_{role}	AP^{S2}_{role}
QPIC (CVPR2021)	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93	58.8	61.0
UPT (CVPR2022)		31.66	25.94	33.36	35.05	29.27	36.77	59.0	64.5
MUREN (CVPR2023)		32.87	28.67	34.12	35.52	30.88	36.91	<u>68.8</u>	<u>71.0</u>
$HybHOI_s$ (CVPR2024)		35.86	32.48	<u>36.86</u>	<u>39.48</u>	36.10	<u>40.49</u>	61.1	66.6
DPAD (AAAI2024)		<u>35.91</u>	<u>35.82</u>	35.94	38.99	<u>39.61</u>	38.80	62.6	64.8
$HOIMamba_s$ (Ours)		**41.51**	**42.75**	**41.15**	**43.41**	**44.77**	**43.01**	**69.2**	**71.1**
CDN (NeurIPS2021)	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38	63.9	65.9
GEN-VLKT (CVPR2022)		34.95	31.18	36.08	38.22	34.36	39.37	63.6	65.9
RmLR (ICCV2023)		37.41	28.81	39.97	38.69	31.27	40.91	64.2	70.2
$HybHOI_m$ (CVPR2024)		36.82	33.99	37.66	<u>40.56</u>	<u>37.02</u>	<u>41.69</u>	62.3	68.2
SCTC (AAAI2024)		<u>39.12</u>	<u>36.09</u>	<u>39.87</u>	-	-	-	<u>68.2</u>	<u>72.5</u>
$HOIMamba_m$ (Ours)		**41.77**	**42.91**	**41.43**	**43.44**	**45.01**	**42.97**	**69.4**	**72.7**
FGAHOI(TPAMI2023)	Swin-Large	37.18	30.71	39.11	38.93	31.93	41.02	60.5	61.2
PViC (ICCV2023)		44.32	44.61	44.24	47.81	48.38	47.64	61.7	68.0
MP-HOI-L (CVPR2024)		44.53	44.48	44.55	-	-	-	-	-
$HybHOI_l$ (CVPR2024)		<u>46.01</u>	<u>46.74</u>	<u>45.80</u>	<u>49.50</u>	<u>50.59</u>	<u>49.18</u>	<u>63.0</u>	<u>68.7</u>
$HOIMamba_l$ (Ours)		**47.66**	**50.01**	**46.96**	**50.44**	**51.88**	**50.01**	**69.9**	**73.1**

Table 1: Comparison of different methods on HICO-DET and V-COCO datasets. **Bold** and <u>underline</u> show the best results and the second best results.

and Hutter 2017) and set the weight decay to 1e-4. The rank r in LoRA is set to 8. The model is trained with 60 epochs and the initial learning rate is reduced to 1e-5 at 50 iterations. To ensure a fair comparison, the data augmentation techniques of the model are consistent with the previous methods (Kim, Jung, and Cho 2023). All experiments are carried out on 4 RTX3090 GPUs with batch size set to 16.

Comparison with State-of-the-Art Methods

We compare the performance with Transformer-based methods including QPIC (Tamura, Ohashi, and Yoshinaga 2021), UPT (Zhang, Campbell, and Gould 2022), MUREN (Kim, Jung, and Cho 2023), HybHOI (Wu et al. 2024), DPAD (Gao et al. 2024), CDN (Zhang et al. 2021), GEN-VLKT (Liao et al. 2022), RmLR (Cao et al. 2023), SCTC (Jiang et al. 2024), FGAHOI (Ma et al. 2023), PViC (Zhang et al. 2023), MP-HOI (Yang et al. 2024a). As shown in Table 1, we set HOIMamba under different backbones to demonstrate the scalability of HOIMamba and to facilitate a fair comparison with previous state-of-the-art methods. Our method significantly outperforms all single-branch, two-branch, and three-branch methods, achieving new state-of-the-art performance under various settings of HICO-DET and V-COCO. Specifically, HOIMamba achieves the full mAP of 41.51, 41.77, and 47.66 across ResNet-50, ResNet-101 and Swin-Large (Liu et al. 2021) feature extractors, which are **15.6%**, **6.8%**, and **3.6%** higher than the previous SOTA method.

Notably, the boost of HOIMamba is very significant in the rare HOIs. Compared to the recent SOTA, three variants of HOIMamba, with HICO-DET default settings, achieve mAP improvements of **+6.93**, **+6.82**, and **+3.27**, respectively. These results validate the superiority of the introduced progressive learning strategy.

Method	Full	Rare	Non-rare
Base	29.07	21.85	31.23
+CEM	$30.10^{(+1.03)}$	$23.39^{(+1.54)}$	$32.11^{(+0.88)}$
+distangling	$32.28^{(+2.18)}$	$28.23^{(+4.84)}$	$33.49^{(+1.38)}$
+prior	$33.07^{(+0.79)}$	$29.58^{(+1.35)}$	$34.12^{(+0.63)}$
+h-o interact	$34.32^{(+1.25)}$	$30.15^{(+0.57)}$	$35.57^{(+1.45)}$
+router	$35.92^{(+1.60)}$	$33.71^{(+3.56)}$	$36.59^{(+1.02)}$
+PL	$40.41^{(\mathbf{+4.49})}$	$41.96^{(\mathbf{+8.25})}$	$39.95^{(\mathbf{+3.36})}$
+DCP	$41.51^{(+1.10)}$	$42.75^{(+0.79)}$	$41.15^{(+1.20)}$

Table 2: Ablation study on overall model framework design.

Ablaion Study

Overall network architecture design. To analyze the effectiveness of different components or strategies within HOIMamba, we use QPIC as the $Base$, as shown in Table 2. The first modification is to replace self-attention and cross-attention in the decoder with CEM block ($+CEM$), which can obtain a 1.54 mAP improvement for rare categories, demonstrating the effectiveness of multi-view adaptive aggregation information. Next, we integrate the benefits of existing two-branch and three-branch branches into our model. We can observe that decoupling the three branches ($+distangling$) leads to a 4.84 rare mAP boost, while providing a prior for interactions via the human and object branches ($+prior$) leads to a significant improvement of +1.35 rare mAP. To demonstrate the need for content exchange between human and object branches, we conduct human and object interaction in the second CEM block ($+h-o\ interact$), and the performance of rare categories is further improved to 30.15 mAP. In addition, we introduce router

(a) <install/repair, traffic light>

(b) <ride/stand on/wear, snowboard>

(c) <paint/inspect, fire hydrant>

(d) <dry/hold/hug/pet, cat>

Figure 4: Visualization of HOI detection results. We mark the correct results in green and the wrong ones in red.

#	φ_1	φ_2	Full	Rare	Non-Rare
A_1	1.0	1.0	31.55	29.55	32.16
A_2	0.75	1.0	34.53	35.57	34.22
A_3	0.50	1.0	38.39	39.15	38.17
A_4	0.25	1.0	38.00	38.25	37.93
A_5	**0.50**	**0.75**	**41.51**	**42.75**	**41.15**
A_6	0.50	0.50	40.76	41.37	40.58

Table 3: Ablation study on the hyperparameters in the CEM.

$(+router)$ to simultaneously realize the implicit gradient separation between different branches and the adaptive enhancement of necessary semantics, improving all categories' performance to 35.92 mAP. To address the lack of interaction gradual learning in existing methods, we introduce a comprehensive progressive learning strategy that combines shallow and deep task-shared semantics to gradually aggregate four-level interaction features from the human and object branches $(+PL)$, obtaining a significant improvement of **+8.25** mAP on rare categories. Finally, we adopt the detection relation context block $(+DCP)$ to pass the necessary semantics in the interaction branch to the two detection branches, which makes a **+1.10** full mAP improvement.

The impact of hyperparameters in the CEM Block. φ_1 is used to control the mixing degree of two inputs within the CEM block, while φ_2 affects the input weight in the parameter-aware scan mechanism. As shown in Table 3, we observe the best HOI detection performance is achieved when φ_1, φ_2 are set to 0.50 and 0.75 respectively.

The impact of the DCP block. As shown in Table 4, by gradually adding the DCP block to the human and object branches, we can observe that the DCP block is needed to facilitate the detection of humans and objects. Introducing the DCP block simultaneously in both detection branches further improves HOI detection performance compared to model $B_1 - B_3$, indicating that detecting context propagation is crucial for identifying more HOI instances.

Analysis of Model Efficiency. We compare the complexity of the proposed HOIMamba with QPIC, AS-Net (Chen et al. 2021) and MUREN under the same backbone (ResNet-50), as shown in Table 5. HOIMamba is less than all methods in the number of parameters, which is even more efficient than single-branch QPIC, showing 36.0% relative improvement of FPS and 15.6% reduction of FLOPS, further proving that our method well absorbs the simplicity and efficiency of the single-branch method.

#	Human	Object	Full	Rare	Non-Rare
B_1	-	-	40.41	41.96	39.95
B_2	✓	-	41.14	42.39	40.77
B_3	-	✓	40.75	42.21	40.31
B_4	✓	✓	**41.51**	**42.75**	**41.15**

Table 4: Ablation study on the DCP Block.

Method	Params	GFLOPs	FPS
QPIC (CVPR2021)	42.35M	36.91	20.04
AS-Net (CVPR2021)	59.14M	52.92	1.63
MUREN (CVPR2023)	69.31M	63.71	1.12
HOIMamba (Ours)	**40.31M**	**31.15**	**27.26**

Table 5: Analysis of efficiency. All models are tested using one RTX 3090 with an input of 640× 640 resolution.

Qualitative Results and Limitations

We offer several qualitative detection results as shown in Figure 4. Specifically, thanks to the well-designed modules and a comprehensive progressive learning strategy, our model recognizes complex interactions in Figure 4 (a) (severe occlusion), Figure 4 (b) (tiny visual features), and Figure 4 (c) (tiny visual features). Meanwhile, HOIMamba also correctly identifies the rare interaction categories ⟨*human, paint, fire hydrant*⟩ in Figure 4 (c) and ⟨*human, dry, cat*⟩ in Figure 4 (d). In addition, we also provide a failure case in Figure 4 (d), our model fails to identify ⟨*human, pet, cat*⟩. This may be due to the limitations of visual representation.

Conclusion

In this work, we propose HOIMamba, a novel HOI detection architecture with a well-designed Mamba-based decoder to mine the benefits of existing methods and enhance the ability to recognize difficult HOI samples. For mining the benefits of existing methods, HOIMamba builds an efficient and effective decoder through cascaded Low-Rank Adaptations (LoRAs), with high efficiency, thorough decoupling of tasks, and good multi-task collaborative learning. For recognizing difficult HOI samples, a Mamba-based comprehensive progressive learning strategy with Cross-enhance Mamba (CEM) blocks and Detection Context Propagation (DCP) blocks is designed to extract more expressive HOI representation. Extensive experiments demonstrate the superiority of the proposed HOIMamba.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62106245, 62476260 and 62436008.

References

Cao, Y.; Tang, Q.; Yang, F.; Su, X.; You, S.; Lu, X.; and Xu, C. 2023. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23492–23503.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 213–229. Springer.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision*, 381–389. IEEE.

Chen, M.; Liao, Y.; Liu, S.; Chen, Z.; Wang, F.; and Qian, C. 2021. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9004–9013.

Dou, S.; Zhou, E.; Liu, Y.; Gao, S.; Zhao, J.; Shen, W.; Zhou, Y.; Xi, Z.; Wang, X.; Fan, X.; et al. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*, 4(7).

Fang, S.; Lin, Z.; Yan, K.; Li, J.; Lin, X.; and Ji, R. 2023. HODN: Disentangling Human-Object Feature for HOI Detection. *IEEE Transactions on Multimedia*.

Gao, C.; Xu, J.; Zou, Y.; and Huang, J.-B. 2020. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, 696–712. Springer.

Gao, J.; Liang, K.; Wei, T.; Chen, W.; Ma, Z.; and Guo, J. 2024. Dual-Prior Augmented Decoding Network for Long Tail Distribution in HOI Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1806–1814.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2024. MambaIR: A Simple Baseline for Image Restoration with State-Space Model. *arXiv preprint arXiv:2402.15648*.

Gupta, S.; and Malik, J. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.

Hu, B.; Liu, J.; Zheng, Y.; Zheng, K.; and Zha, Z.-J. 2024. Exert Diversity and Mitigate Bias: Domain Generalizable Person Re-identification with a Comprehensive Benchmark. *International Journal of Computer Vision*, 1–27.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiang, W.; Ren, W.; Tian, J.; Qu, L.; Wang, Z.; and Liu, H. 2024. Exploring Self-and Cross-Triplet Correlations for Human-Object Interaction Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2543–2551.

Kim, B.; Lee, J.; Kang, J.; Kim, E.-S.; and Kim, H. J. 2021. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 74–83.

Kim, S.; Jung, D.; and Cho, M. 2023. Relational context learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2925–2934.

Liao, Y.; Zhang, A.; Lu, M.; Wang, Y.; Li, X.; and Liu, S. 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20123–20132.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

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 *European Conference on Computer Vision*, 740–755. Springer.

Liu, J.; Zha, Z.-J.; Chen, D.; Hong, R.; and Wang, M. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7202–7211.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, S.; Wang, Y.; Wang, S.; and Wei, Y. 2023. Fgahoi: Fine-grained anchors for human-object interaction detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ning, S.; Qiu, L.; Liu, Y.; and He, X. 2023. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23507–23517.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666.

Tamura, M.; Ohashi, H.; and Yoshinaga, T. 2021. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10410–10419.

Wan, B.; Zhou, D.; Liu, Y.; Li, R.; and He, X. 2019. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9469–9478.

Wu, E. Z.; Li, Y.; Wang, Y.; and Wang, S. 2024. Exploring Pose-Aware Human-Object Interaction via Hybrid Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17815–17825.

Yang, J.; Li, B.; Zeng, A.; Zhang, L.; and Zhang, R. 2024a. Open-World Human-Object Interaction Detection via Multi-modal Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16954–16964.

Yang, Y.; Zhai, W.; Luo, H.; Cao, Y.; Luo, J.; and Zha, Z.-J. 2023. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10905–10915.

Yang, Y.; Zhai, W.; Luo, H.; Cao, Y.; and Zha, Z.-J. 2024b. LEMON: Learning 3D Human-Object Interaction Relation from 2D Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16284–16295.

Zhang, A.; Liao, Y.; Liu, S.; Lu, M.; Wang, Y.; Gao, C.; and Li, X. 2021. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34: 17209–17220.

Zhang, F. Z.; Campbell, D.; and Gould, S. 2022. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20104–20112.

Zhang, F. Z.; Yuan, Y.; Campbell, D.; Zhong, Z.; and Gould, S. 2023. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10411–10421.

Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; and Ouyang, W. 2024. RS-Mamba for Large Remote Sensing Image Dense Prediction. *arXiv preprint arXiv:2404.02668*.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.