Category-Aware Transformer Network for Better Human-Object Interaction Detection

Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, Xu Zou[*]

National Key Laboratory of Science and Technology on Multispectral Information Processing,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
{dongleizhen36, lizm, xukunlun, zoux}@hust.edu.cn

Abstract

Human-Object Interactions (HOI) detection, which aims to localize a human and a relevant object while recognizing their interaction, is crucial for understanding a still image. Recently, tranformer-based models have significantly advanced the progress of HOI detection. However, the capability of these models has not been fully explored since the Object Query of the model is always simply initialized as just zeros, which would affect the performance. In this paper, we try to study the issue of promoting transformer-based HOI detectors by initializing the Object Query with category-aware semantic information. To this end, we innovatively propose the Category-Aware Transformer Network (CATN). Specifically, the Object Query would be initialized via category priors represented by an external object detection model to yield a better performance. Moreover, such category priors can be further used for enhancing the representation ability of features via the attention mechanism. We have firstly verified our idea via the Oracle experiment by initializing the Object Query with the groundtruth category information. And then extensive experiments have been conducted to show that a HOI detection model equipped with our idea outperforms the baseline by a large margin to achieve a new state-of-the-art result.

1. Introduction

Human-Object Interaction (HOI) detection, serving as a fundamental task for high-level computer vision tasks, e.g. image captioning, visual grounding, visual question answer, etc., has attracted enormous attention in recent years. Given an image, HOI detection aims to localize the pair of human and object instances and recognize the interaction between them. A human-object interaction could be defined as the <human, object, verb> triplet.

Many two- or one-stage methods [2, 6, 7, 12, 13, 17, 18, 20, 26, 34, 35, 38] have significantly advanced the process of HOI detection, while transformer-based methods [3, 14, 19, 30, 36, 39] have been remarkably proposed recently and achieved the new state-of-the-art result.

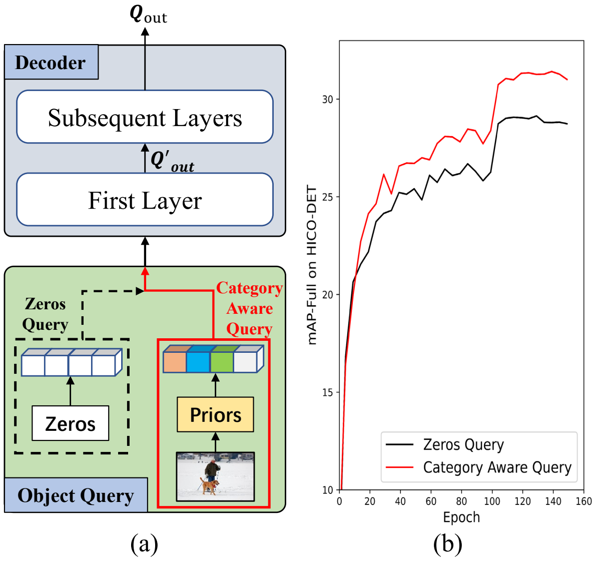


Figure 1. **Black-Dotted-Box** in (a) and **Black-Curve** in (b): The Object Query of the first decoder layer is simply initialized as just zeros in all previous transformer-based HOI detection methods. **Red-Box** in (a) and **Red-Curve** in (b): Our idea is to take the category-aware information as the Object Query initialization, where the training curve indicates that our method significantly promotes mAP from 29.07 to 31.03 on the HICO-DET dataset.

Thanks to the self-attention and cross-attention mechanisms, Transformer [32] has a better capability of capturing long-range dependence between different instances, which is especially suitable for the HOI detection. HOTR [14] and AS-Net [3] utilize two parallel branches with transformer-decoders for performing instance detection and interaction classification respectively. Motivated by DETR [1], HOI-Transformer [39] and QPIC [30] adopt one transformer-decoder with several sequential layers and automatically group the different types of predictions from one query into an HOI triplet in an end-to-end manner.

[*] Corresponding author.

Though these transformer-based methods have greatly promoted the community by improving the performance and efficiency without complex grouping strategies, there is a common issue with the Object Query[†] regardless of the differences in these methods. Specifically, the Object Query of the first decoder layer of these methods is always simply initialized as zeros since there is no previous layer for feeding semantic features (shown as Black-Dotted-Box in Figure 1(a)). The capability of these models has not been fully explored due to the simple initialization of the Object Query, which would affect the performance. Meanwhile, multi-modal information, including spatial [2], posture [5], and language [37], has been indicated to be beneficial for two-stage HOI detection models. Thus, one question remains: **how semantic information promotes a transformer-based HOI detection model?** In this paper, we try to study the issue of elevating transformer-based HOI detectors by initializing the Object Query with category-aware semantic information.

To this end, we present the Category-Aware Transformer Network (CATN), consisting of two modules: the Category Aware Module (CAM) and the Category-Level Attention Module (CLAM). CAM can obtain category priors which is then applied to the initialization of the Object Query. Specifically, we use an external object detector and design a select strategy to get the categories contained in the image. After that, these categories would be transferred to corresponding word embeddings as final category priors. Moreover, these priors could be further used for enhancing the representation ability of features via the proposed CLAM.

We first show that category-aware semantic information can indeed promote a transformer-based HOI detection model by the Oracle Experiment where the category priors are generated from the ground truth. Then we evaluate the effectiveness of our proposed CATN on two widely used HOI benchmarks: HICO-DET and V-COCO datasets. The contributions of our work could be summarized as:

- We reveal that a transformer-based HOI model can be further improved by taking category-aware semantic information as the initialization of the Object Query to achieve a new state-of-the-art result.
- We present the Category-Aware Transformer Network (CATN), which obtains two modules: CAM for generating category priors of an image and CLAM for enhancing the representation ability of the features.
- Extensive experiments, involving discussions of different initialization types and where to leverage the semantic information, have been conducted to demonstrate the effectiveness of the proposed idea.

2. Related Works

Many remarkable methods have advanced the progress of HOI detection, which could be simply categorized into Two-, one-stage methods, and transformer-based methods.

Two-stage methods. Two-stage methods usually utilize a pre-trained object detector to generate human and object proposals in the first stage and then adopt an independent module to infer the multi-label interactions of each human-object pair in the second stage. HO-RCNN [2] firstly presents a multi-stream architecture. iCAN [7] proposes an Instance-Centric Attention to aggregate the context feature of humans and objects. In order to obtain accurate interactions, some extra information, e.g. human posture [17, 33] and language knowledge [37], have been introduced into HOI detection. To better model the spatial relationship between the human and object, some GNN-based methods [6, 26, 31, 38] are sequentially proposed and improve the performance. Two-stage methods generally suffer from inefficiency due to the separate architecture, where all possible pairs of human-object proposals are predicted one after the other and the cropped features generated from the object detector maybe not suitable for interaction classification in the second stage.

One-stage Methods. One-stage methods are proposed to deal with the problems of high computational cost and feature mismatching appearing in two-stage methods. PPDM [20] and IPNet [34] address the task of HOI as a keypoint detection problem by regarding the interaction point as the mid of human-object centers. Based on the feature at the midpoint, the interactions between the human and object are predicted in a one-stage manner. Meanwhile, UnionNet [13] provides another alteration to perform HOI detection in a one-stage manner, which treats the union box of human and object bounding-box as the region of each HOI triplet. UnionNet conduct an extra branch to predict the union box and group the final HOI triplet based on IoUs. Despite great improvement in efficiency, the performance of existing one-stage methods is limited by complex handcrafted grouping strategies to group object detection results and the interaction predictions into final HOI triplets.

Transformer-based Methods. Recently, transformer [32], with a good capability of capturing the long-range dependency, has been introduced to the HOI detection and brings a significant performance improvement. HOTR [14] and AS-Net [3] combine the advantages of both one-stage method and transformer, and utilize two parallel decoders to predict human-object proposals and interactions respectively. Apart from the above methods, HOI-Transformer [39] and QPIC [30] extend DETR [1] to the HOI detection, which directly defines the predictions from a query as the HOI triplet without the complex grouping strategy.

Although significant performance is obtained by these

[†] The Object Query is one input of the transformer-decoder and contains N_q object queries without query positional embedding in this paper.

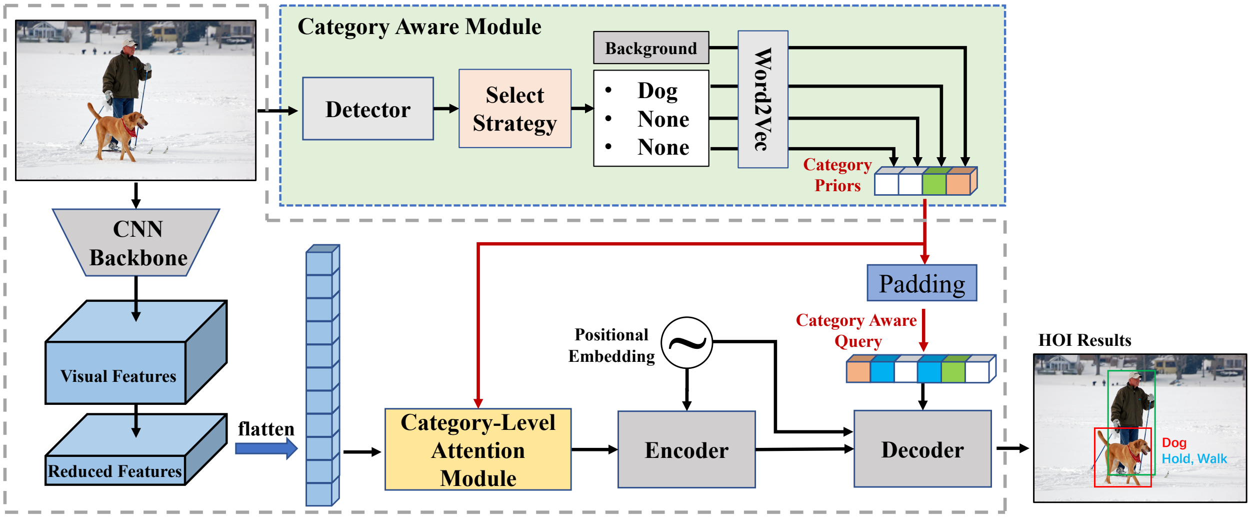


Figure 2. Overall architecture of our proposed CATN. Compared with the previous, our method contains two main components: Category Aware Module (CAM) and Category-Level Attention Module (CLAM). We propose CAM which uses an external object detector to obtain category priors of the image and the priors are then applied for initializing the Object Query. Moreover, such priors can be further used in the CLAM for enhancing the representation ability of features.

transformer-based methods, they have a common issue that the Object Query is initialized with zeros as illustrated in Figure 1(a). In this paper, we study the issue of how to promote a transformer-based HOI detector by initializing the Object Query with category-aware semantic information.

Category Information and HOI Detection. Category Information is a kind of semantic information indeed, which represents the object categories in an image. The effectiveness of such information has been demonstrated in several domains. Different from part-of-speech category for image captioning [15] or the category shape for 3D-Reconstruction [29], we studied object category information since the instance has category-aware relation in HOI detection, e.g. person-eat-apple, person-ride-bike, etc.

3. Approach

3.1. Overview

In this section, we present our Category-Aware Transformer Network (CATN), trying to improve the performance of transformer-based HOI detectors with category priors. Firstly, we start with the overall architecture of our proposed CATN. Secondly, we detailedly introduce the Category Aware Module (CAM) to extract category priors of an image, which then are applied to the initialization of the Object Query of the first decoder layer. Moreover, we propose the Category-Level Attention Module (CLAM) to enhance the capability of features with such priors. Finally, we modify the matching cost, used in bipartite matching, for better matching between the ground-truths and N_q predictions.

3.2. CATN Architecture

The overall pipeline of our CATN is illustrated in Figure 2, which is similar to previous transformer-based methods except for additional proposed CAM and CLAM.

Backbone. Given an RGB image, we firstly adopt a CNN-based backbone to extract a visual feature map denoted as $I_c \in \mathbb{R}^{D_c \times h \times w}$. Then a convolution layer with a kernel size of 1×1 is utilized to reduce the channel dimension from D_c to D_d, where D_c, D_d are 2,048 and 256 by default. Then the visual feature map is flattened and denoted as $I_{visual} \in \mathbb{R}^{D_d \times hw}$. After that, we adopt the CLAM to enhance the features from CNN with category priors and denote the output feature map as $I_{CLAM} \in \mathbb{R}^{D_d \times hw}$.

Encoder. The transformer encoder aims to improve the capability of capturing long-range dependence. It is a stack of multiple encoder layers, where each layer mainly consists of a self-attention layer and a feed-forward (FFN) layer. To make the flatten features spatially aware, a fixed Spatial Positional Encoding, denoted as $P_S \in \mathbb{R}^{D_d \times hw}$, is conducted and fed into each encoder layer with the features. The calculation of the transformer encoder could be expressed as:

$$I_{enc} = f_{enc \times N_{enc}}(I_{CLAM}, P_S) \tag{1}$$

where f_{enc} indicates the function of one encoder layer, N_{enc} is the number of stacked layers, and $I_{enc} \in \mathbb{R}^{D_d \times hw}$ is the output feature and then fed into the following decoder.

Decoder. The transformer decoder aims to transform a set of object queries $Q_{zeros} \in \mathbb{R}^{N_q \times D_d}$ (with query positional embedding $P_Q \in \mathbb{R}^{N_q \times D_d}$ whose parameters are learnable) to another set of output queries $Q_{out} \in \mathbb{R}^{N_q \times D_d}$. It is also a stack of decoder layers. Apart from selt-attention and FFN, each decoder layer contains an additional cross-attention layer, which is used to aggregate the features I_{enc} output from encoder into N_q queries.

In our CATN, Q_{zeros} is replaced with Category Aware Query (CAQ), denoted as $Q_{CA} \in \mathbb{R}^{N_q \times D_d}$, which is generated via category priors. The calculation of the transformer decoder could be expressed as:

$$Q_{out} = f_{dec \times N_{dec}}(Q_{CA}, P_Q, I_{enc}, P_S) \tag{2}$$

where f_{dec} indicates the function of one decoder layer and N_{dec} is the number of stacked decoder layers.

Prediction Head. In our experiments, an HOI triplet consists of four elements: the human bounding box, the object bounding box, the object category with its confidence, and multiple verb categories with their confidence. Based on the above definition, four feed-forward networks (FFNs) are conducted on each output query as follows:

$$\begin{cases} b_h^i = \sigma(f_{h,b}(Q_{out}^i)) \\ b_o^i = \sigma(f_{o,b}(Q_{out}^i)) \\ c_o^i = \varsigma(f_{o,c}(Q_{out}^i)) \\ c_v^i = \sigma(f_{v,c}(Q_{out}^i)) \end{cases} \tag{3}$$

where i indicates the index of outputting queries and the ground-truths, and σ, ς are the sigmoid and softmax functions respectively.

3.3. Category Aware Module

As mentioned above, the Object Query of the first decoder layer is simply initialized with zeros since there is no last layer where we argue this may affect the performance. In this section, we detailedly introduce the Category Aware Module (CAM) to extract the category priors of an image which then are used for Category Aware Query and CLAM.

The Blue-Dotted-Box in Figure 2 describes the proposed CAM. Given an image, we firstly utilize an external object detector, e.g. Faster-RCNN, to perform object detection and only reserve the results with confidence scores higher than the detection threshold T_{det}. Since we focus on studying the effect of category-aware semantic information on HOI detectors and avoid the influence of other factors, we directly discard the bounding box of each prediction and only utilize the category with its confidence score.

Select Strategy. Based on their categories, the rest results can be divided into different sets $\Omega = \{\Omega_1, \Omega_2, ..., \Omega_K\}$, where K is the total number of categories in the dataset and Ω_i represents a set of detection results whose category is the i-th category denoted as c_i. After that, we calculate the confidence score as follows:

$$S_{c_i} = \begin{cases} max(\Omega_i) + \frac{|\Omega_i|}{2} \times mean(\Omega_i), & |\Omega_i| \neq 0 \\ 0, & |\Omega_i| = 0 \end{cases} \tag{4}$$

where S_{c_i} represents the probability of category c_i contained in the image and $|.|$ indicates the number of the set.

With these statistics, we firstly select a threshold T_{can} for a set of candidate categories $\Omega_{can} = \{c_i | \sum_{i=1}^{K} S_{c_i} \geq T_{can}\}$ and re-rank them based on S_{c_i}. Then $Top^{(N_c-1)}$ categories from Ω_{can} with a fixed category (named as 'background') are set as the prior categories of an image, where the 'background' is used as the placeholder for matching if no relevant instance is obtained by the detector in CAM and the detail is discussed in Section 3.5. Note that the rest category will be filled with 'None' if the number of categories in Ω_{can} is lower than $N_c - 1$. We denote the final prior categories of the image as $C^* = \{c_i | c_i \in C_{can} \cup None\}_i^{N_c}$.

Category Priors. We transform prior categories of an image to the word embedding vectors which could be used in the following module. To this end, we utilize a pre-trained word2vector model, e.g. fastText [24], to generate the category priors of an image.

$$E_{prior} = \{f_{FC}(f_{w2v}(c_i)) | c_i \in C^*\} \tag{5}$$

where $f_{w2v}(c_i)$ is to obtain the embedding vector of c_i category and f_{FC} is a fully connected layer to adjust the dimension of the embedding. Especially, the embeddings of all object categories are calculated beforehand and saved locally. Regardless of training or inference, there is only a slight increase in computation cost due to the fully connected layer. In addition, we also evaluate several different word2vector models and the experimental results are shown in the later section.

Category Aware Query. An image may contain more than one HOI triplet with the same category and the number of prior categories N_c is usually much smaller than the number of queries N_q. Thus, we generate $Q_{CA} \in R^{N_q \times D_d}$ by simply repeating the E_{prior} vectors $\frac{N_q}{N_c}$ times as follows.

$$Q_{CA} = Repeat(E_{prior}, N_q, N_c) \tag{6}$$

Finally, we use Q_{CA} as initial values of the Object Query.

3.4. Category-Level Attention Module

For maximizing the capability of category information, we also propose an attention mechanism, named as Category-Level Attention Module (CLAM), to enhance the representation ability of features output from backbone. As illustrated in Figure 3, to clearly describe the entire workflow, we take the visual feature in one location as an example of this module and denote the feature as $X_{visual} \in \mathbb{R}^{1 \times D_d}$, while features work consistently.

The visual feature X_{visual} is firstly projected to another D_d-dimensional vector, denoted as $\hat{X}_{visual} \in \mathbb{R}^{1 \times D_d}$, via a Muti-Layer Perception (MLP). The MLP contains an FC layer without BatchNorm and ReLU and is used to transform the feature from visual space to word space. Then we

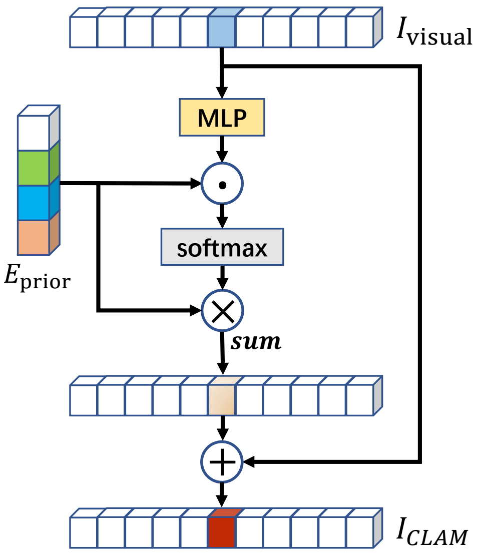


Figure 3. The pipeline of our proposed Category-Level Attention Module (CLAM). Each cuboid indicates a feature vector with the shape of $1 \times D_d$. $\cdot$, $\times$, $+$ mean dot-product, multiplication and element-wise addition respectively.

measure the similarity between the vector and category embedding by dot product operation, and a softmax function is followed to normalize the similarity values of all categories.

$$W_{att} = Softmax(MLP(X_{visual}) \cdot E_{prior}^{T}) \tag{7}$$

where $W_{att} \in \mathbb{R}^{1 \times N_c}$ represents the attention weights of prior categories on the feature and '$\cdot$' means Dot-Product. Specifically, the weight in a category is high if the feature contains rich information related to the category. To make the features category aware, we aggregate all word embeddings of prior categories with corresponding weights into one vector $X_{word} \in \mathbb{R}^{1 \times D_d}$ and then add the vector back to the original visual feature X_{visual}. The feature aggregation can be written as:

$$X_{clam} = X_{app} + \sum_{j=1}^{N_c} w_{att}^{j} \times E_{prior}^{j} \tag{8}$$

where w_{att}^{j} is a value and indicates the attention weight of j-th category belonging to the prior categories, E_{prior}^{j} is the embedding of j-th category, '+' represents element-wise addition, and '$\times$' mean the multiplication between the scalar w_{att}^{j} and each element of the vector E_{prior}^{j} respectively.

Our proposed CLAM has several advantages. Compared to instance-level attention mechanism [7], ours is category-level and has lower requirements on the accuracy of bounding boxes. Meanwhile, we add the weighted average word embedding with category information to the originally visual features. Therefore, the aggregated features, output from our CLAN, not only have the capability of aware visual information but also are aware of category information.

3.5. Matching Cost & Training Loss

For training transformer-based models [1] with a set of prediction results, the bipartite-matching algorithm is publicly used to automatically match a ground-truth with at most one prediction, which would suppress the problem of redundant predictions. To this end, two types of losses are introduced below.

Modified Matching Cost. Matching cost is conducted to measure the similarity between the ground truth and an HOI prediction and assign the label of each query whether a positive or a negative. Firstly, we calculate the matching cost $H \in \mathbb{R}^{N_{gt} \times N_q}$ following Formula 1 in [30], where N_{gt} is the number of ground-truth HOI triplet and $H_{i,j}$ indicates the matching cost between i-th ground-truth and j-th prediction generated from j-th output query. Then, we modify the matching cost by $\hat{H}_{i,j} = H_{i,j} + Cost_{i,j}$ and the external cost $Cost_{i,j}$ is defined as follows:

$$Cost_{i,j} = \begin{cases} 0, & C(q_j) = C(GT_i) \\ v, & C(q_j) = \text{``}Background''\\ 2v, & C(q_j) = \text{``}None'' \\ 2v, & Else \end{cases} \tag{9}$$

where C represents the corresponding object category, q_j is the j-th query of Q_{CA}, GT_i means the i-th ground-truth triplet in the image. The $Cost_{i,j}$ is used to make the matching cost $H_{i,j}$ higher when the object categories of j-th query and i-th ground-truth are different. Meanwhile, the experimental results show that there is no difference when the value is higher than a threshold. Thus we empirically set v as 500. Finally, we utilize the Hungarian Algorithm [16] to perform the optimal assignment $\hat{\omega} = argmin_{\omega \in \Omega_{N_q}} \sum_{i=1}^{N_q} \hat{H}_{i,\omega(i)}$, where only N_{gt} predictions in $\hat{\omega}$ are set as positive and the rest are negative.

With the above modifications, a ground-truth will match the query where their object category are the same. In addition, if the object category of a ground truth is not included in prior categories, the ground truth will preferentially match the query whose prior category is "background". Thus the modified cost shrinks the matching space between the ground truth and the predictions.

Training Loss. Based on the above label assignment, the training loss is calculated to optimize the parameters of our CATN model. We directly adopt equations 6∼10 in [30] and keep the weights consistent, which reveals that the performance improvement is obtained by our proposed category priors, not hyper-parameters.

4. Experiments

4.1. Datasets & Metrics

Datasets. We conduct experiments on two widely used datasets to verify the effectiveness of our model. **V-**

COCO [9], a subset of COCO [22] dataset, consists of 2,533 training images, 2,867 valuation images, and 4,946 test images respectively. There are 16,199 human instances and each instance has a set of binary labels for 29 different actions. **HICO-DET** [2] is the largest dataset in HOI detection. There are 38,118 training images and 9,658 test images respectively with totally more than 150k HOI annotations. It has 600 hoi categories (Full) with 117 verb categories and 80 object categories, which can be further divided into 118 categories (Rare) and 462 categories (Non-Rare) based on the number of instances in the training set.

Evaluation Metrics. Following the standard rule [9], we use the commonly used role mean average precision (mAP) to evaluate the model performance for both benchmarks. An HOI prediction is regarded as a true positive if the categories of the object and verbs are correct, and the predicted bounding boxes of the human and object are localized accurately where the IoUs are greater than 0.5 with the corresponding ground truth.

4.2. Implementation Details

We conduct our experiments with the publicly available PyTorch framework [25].

For the external detector used in CAM to obtain category priors, we adopt Faster-RCNN-FPN [21, 28] with ResNet-50 [10] as the backbone to perform object detection. For better performance, we use COCO pre-trained weights and then fine-tune the model on both HICO-DET and V-COCO datasets. During training, we drop the weight of regressing Bbox in loss cost from 1.0 to 0.2 and keep the other hyper-parameters consistent as default. Meanwhile, the human category is discarded since the number of humans is dominant. We set the batch size to 4 and use SGD as the optimizer with a learning rate of 0.01, a weight decay of 0.0001, and a momentum of 0.9. We train the model for 12 epochs with twice the learning rate decay at epoch 8, 11 by 10 times respectively. The detection threshold T_{det} is set to 0.15. The prior threshold T_{can}, used for category priors, is set to 0.3 and 0.4 for HICO-DET and V-COCO respectively. The number N_c is set to 4 and 5 for two datasets respectively. To obtain better category priors, we adopt some commonly used augmentation strategies, including random scales, random flip, color jittering, and random corp augmentation.

For our CATN, ResNet-50 is used as the backbone, the number of encoder and decoder layers are both set to 6 and the number of Object Query N_q is set to 100. We initialize the network with parameters of DETR [1] pre-trained on the COCO dataset. During training, we set the batch size to 16, the backbone's learning rate to 1e-5, the transformer's learning rate to 1e-4, and weight decay to 1e-4. The model is trained for 150 epochs totally on both datasets with once learning rate decreased by 10 times at epoch 100.

Method	Query	Full	Rare	Non-Rare
baseline	Zeros	29.07	21.85	31.23
Ours	CAQ*	37.17	31.65	38.81
Improvement		(8.10 ↑)	(9.80 ↑)	(7.58 ↑)

Table 1. Oracle experiment on HICO-DET dataset. Zeros and CAQ represent that the Object Query is initialized with zero-values or category priors respectively. * indicates such category priors generated from the ground truth. The performance is tremendously promoted once category priors are adopted for initializing the Object Query. This phenomenon directly indicates the rationality of introducing category priors.

Following DETR, scale augmentation, scaling the input image such that the shortest side is at least 480 and at most 800 pixels while the longest at most 1333, is adopted for better performance in training.

Note that the category embedding is generated by fastText [24] and the "baseline" indicates the QPIC [30] with ResNet-50 [10] if there are no additional comments.

4.3. Oracle Experiment

To verify the effectiveness of our idea that the performance of transformer-based HOI detectors could be further improved by initializing the Object Query with category-aware semantic information, we firstly conduct the oracle experiment where the category priors of an image are simply generated from the ground truth.

Table 1 illustrates the experimental results on HICO-DET. In this experiment, we select QPIC as the baseline and only apply such priors to the Object Query without the proposed CLAM. Compared with the baseline, our method achieves a great performance improvement on all three default settings. With such category priors, the 'Full' performance is improved from 29.09 to 37.17 with a 27.8% relative performance gain and especially the 'Rare' performance is improved from 21.85 to 31.65 with a 44.8% relative performance gain. This simple experiment with great performance gain verifies the effectiveness of our idea and supports subsequent detailed experiments.

4.4. Comparison to the State-of-The-Art

In this section, we use the proposed CAM to obtain category priors of an image and compare our proposed CATN with other state-of-the-art methods on two public benchmarks. **HICO-DET.** To verify the effectiveness of our proposed idea, we adopt several different word2vector models including fastText [24], BERT [4] and CLIP [27], to obtain the category-aware semantic information and conducts the experiments on HICO-DET dataset. As shown in Table 2, our proposed method obtains the significant performance improvement on both "Default" and "Known-Object" evaluation modes. Especially, the experiment with BERT [4] has achieved the new state-of-the-art result, which promotes

Methods	Backbone	Detector	Default			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
Two-Stage Methods								
HO-RCNN [2]	CaffeNet	C	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [8]	R50-FPN	C	9.94	7.16	10.77	-	-	-
GPNN [26]	R101	C	13.11	9.34	14.23	-	-	-
iCAN [7]	R50	C	14.84	10.45	16.15	16.26	11.33	17.73
PMFNet [33]	R50-FPN	C	17.46	15.65	18.00	20.34	17.47	21.20
VSGNet [31]	R152	C	19.80	14.63	20.87	-	-	-
PDNet [37]	R152	C	20.81	15.90	22.28	24.78	18.88	26.54
FCMNet [23]	R50	C	20.41	17.34	21.56	22.04	18.97	23.12
PastaNet [17]	R50	C	22.65	21.17	23.09	24.53	23.00	24.99
VCL [11]	R101	H	23.63	17.21	25.55	25.98	19.12	28.03
DRG [6]	R50-FPN	H	24.53	19.47	26.04	27.98	23.11	29.43
One-Stage Methods								
UnionDet [13]	R50-FPN	H	17.58	11.52	19.33	19.76	14.68	21.27
IPNet [34]	HG-104	C	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [20]	HG-104	H	21.73	13.78	24.10	24.58	16.65	26.84
Transformer-based Methods								
HOI Transformer [39]	R50	-	23.46	16.91	25.41	26.15	19.24	28.22
HOTR [14]	R50	-	25.10	17.34	27.42	-	-	-
AS-Net [3]	R50	-	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [30]	R50	-	29.07	21.85	31.23	31.68	24.14	33.93
Ours								
CATN (with fastText [24])	R50	H	31.62	24.28	33.79	33.53	26.53	35.92
CATN (with BERT [4])	R50	H	**31.86**	**25.15**	**33.84**	**34.44**	**27.69**	**36.45**
CATN (with CLIP [27])	R50	H	31.71	24.82	33.77	33.96	26.37	36.23

Table 2. Comparison against state-of-the-art methods on HICO-DET dataset. For Detector, C means that the detector is trained on COCO dataset, while H means that the detector is then fine-tuned on HICO-DET dataset. 'Default' and 'Known Object' are two evaluation modes following the standard rule. "fastText", "BERT", "CLIP" means that the embeddings of prior categories are obtained from these pre-trained word2vector models. The BEST and the SECOND BEST performances are highlighted in **bold** and underlined respectively. Our proposed CATN outperforms the previous by a large margin to achieve new state-of-the-art results on both evaluation modes.

Methods	Backbone	AProle
VCL [11]	R50-FPN	48.3
DRG [6]	R50-FPN	51.0
PDNet [37]	R152	52.6
UnionBox [13]	R50-FPN	47.5
IPNet [34]	HG-104	51.0
HOI Transformer [39]	R50	52.9
HOTR [14]	R50	55.2
AS-Net [3]	R50	53.9
QPIC [30]	R50	58.8
CATN (with fastText [24])	R50	**60.1**

Table 3. Comparison against state-of-the-art methods on V-COCO dataset. The BEST performances are high-lighted in **bold**. Ours also outperforms others to achieve a new state-of-the-art result.

the mAP-full from 29.07 to 31.86 in Default mode and from 31.74 to 34.44 in Known-Object mode. **V-COCO.** We also evaluate our proposed CATN on V-COCO dataset. A similar performance gain is obtained as shown in Tabel 3. Compared with previous methods, our method also achieves a new state-of-the-art result. With the embeddings generated by fastText [24], we reach an AP-role of 60.1, which obtains 1.3 points performance gain than the second-best method.

4.5. Ablations Study

The effectiveness of each component in our CATN. In order to make a clearer study of the impact of each component on the overall performance, supplementary ablation experiments are conducted on the HICO-DET dataset. The results in Default evaluation mode are shown in Table 4. Initializing the Object Query with category-aware semantic information instead of just zeros [30] can effectively improve mAP from 29.07 to 30.82, which indicates the superiority of our main idea on HOI detection. Modifying the matching cost can also promote mAP to 31.03 with a gain of 0.21 mAP. Illustrated as line 4, the performance could be further improved from 31.03 to 31.62 when our proposed

	Method	CAQ	MMC	CLAM	mAP
1	baseline	-	-	-	29.07
2	CATN	✓			30.82
3		✓	✓		31.03
4		✓	✓	✓	**31.62**

Table 4. Ablation studies on the effectiveness of each module in our CATN on HICO-DET dataset. ✓ represents the component is used. "CAQ" means the Object Query is initialized with category-aware semantic information. "MMC" indicates our modified matching cost. "CLAM" represents the proposed Category-Level Attention Module.

CLAM enhances the representation ability of features via the category-aware semantic information. Moreover, we visualize an example of the attention map in the supplementary file to demonstrate the effectiveness of our CLAM.

4.6. Discussion

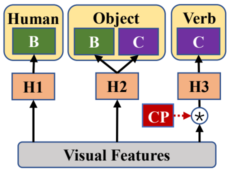
The impacts of where to leverage the category priors. To verify how the category-aware semantic information better promotes the HOI detection model, we design experiments to leverage category priors in another location. As Figure 4, the category priors are introduced in prediction heads. Before predicting the categories of interaction, we combine the visual feature and the category prior by different operations (add and concatenate). Experimental results indicate that taking category-aware semantic information as the Object Query initialization achieves better performance than using the information as complementary features.

The impacts of different initial types. Table 5 presents comparisons to different types of query initialization, including "Zeros", "Random Values (following the Uniform or Gaussian distribution)" and "Category-Aware Semantic Information". Models of the Object Query initialized with 3 different category-aware semantic information consistently achieve better performance than other initial types.

Hyper-parameters in CAM. Figure 5 illustrates the variance by several hyper-parameters, including N_c, T_{det}, and T_{can}, in CAM. To clearly study the impacts1 of them on the quality of category priors, we calculate the recall and precision metrics of the prior categories in image level not instance level. In other words, we only care if a object category could be detected, not the amount and location. We change one parameter in turn and keep others consistent. We achieve the best performance where $N_c = 3$, $T_{det} = 0.15$, and $T_{can} = 0.30$, due to a better trade-off between the recall and precision.

5. Conclusion

In this paper, we explore the issue of promoting a transformer-based HOI model by initializing the Object Query with category-aware semantic information. We propose the Category-Aware Transformer Network (CATN), which obtains two modules: CAM for generating category priors of an image and CLAM for enhancing the representation ability of the features. Extensive experiments, involving discussions of different initialization types and where to leverage the semantic information, have been conducted to demonstrate the effectiveness of the proposed idea. With the category priors, our method achieves new state-of-the-art results on both V-COCO and HICO-DET datasets.



(a)

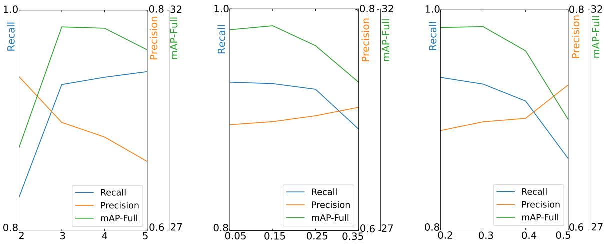
	Method		mAP
1	baseline	-	29.07
2	Head	Add	29.24
3		Concat	29.08
4	**CATN(Ours)**	CAQ	**31.62**

(b)

Figure 4. The impacts of where to leverage the category priors. "H", "B", "C" are prediction heads, bounding boxes, and categories respectively. Similar to [6], Figure (a) and "Head" in Tabel (b) indicate our experiments of introducing category priors (CP) into the verb prediction head. Results from (b) indicate that taking such semantic information as the Object Query initialization (shown as Figure 1.a) achieves a significant performance gain than into the prediction head (Row.4 vs. Row2/3).

	Method	Value	Full	Rare	Non-Rare
1	Zeros	Zero	29.07	21.85	31.23
2	Rondom	Uniform	29.70	23.53	31.53
3		Gaussian	29.60	22.42	31.73
4	CAQ	fastText [24]	31.03	23.97	33.12
5		BERT [4]	31.28	24.89	33.14
6		CLIP [27]	31.23	24.82	33.10

Table 5. The impacts of different initial types. CLAM is not used due to the need of category priors. Models of the Object Query initialized with 3 different category-aware semantic information consistently achieve better performance than other initial types.



(a) N_c. (b) T_{det}. (c) T_{can}.

Figure 5. The impacts of different settings of hyper-parameters, including N_c, T_{det}, T_{can}, in CAM.

6. Acknowledge

This work is supported by the National Natural Science Foundation of China (NSFC) grant 62176100, the Central Guidance on Local Science and Technology Development Fund of Hubei Province grant 2021BEE056.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 5, 6

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision*, pages 381–389. IEEE, 2018. 1, 2, 6, 7

[3] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 1, 2, 7

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6, 7, 8

[5] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision*, pages 51–67, 2018. 2

[6] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision*, pages 696–712. Springer, 2020. 1, 2, 7, 8

[7] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 1, 2, 5, 7

[8] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 7

[9] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[11] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision*, pages 584–600. Springer, 2020. 7

[12] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 1

[13] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *Proceedings of the European Conference on Computer Vision*, pages 498–514. Springer, 2020. 1, 2, 7

[14] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 1, 2, 7

[15] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2019. 3

[16] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[17] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 1, 2, 7

[18] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 1

[19] Zhimin Li, Cheng Zou, Yu Zhao, Boxun Li, and Sheng Zhong. Improving human-object interaction detection via phrase learning and label composition. *arXiv preprint arXiv:2112.07383*, 2021. 1

[20] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 1, 2, 7

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 6

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6

[23] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *Proceedings of the European Conference on Computer Vision*, pages 248–265. Springer, 2020. 7

[24] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*, 2017. 4, 6, 7, 8

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6

[26] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by

graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision*, pages 401–417, 2018. 1, 2, 7

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 6, 7, 8

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 6

[29] Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, et al. Frodo: From detections to 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14720–14729, 2020. 3

[30] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. 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*, pages 10410–10419, 2021. 1, 2, 5, 6, 7

[31] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13617–13626, 2020. 2, 7

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2

[33] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 2, 7

[34] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 1, 2, 7

[35] Kunlun Xu, Zhimin Li, Zhijun Zhang, Leizhen Dong, Wenhui Xu, Luxin Yan, Sheng Zhong, and Xu Zou. Effective actor-centric human-object interaction detection. *Image and Vision Computing*, page 104422, 2022. 1

[36] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[37] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human–object interaction detection. *International Journal of Computer Vision*, 129(6):1910–1929, 2021. 2, 7

[38] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 843–851, 2019. 1, 2

[39] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 1, 2, 7