Group-HOI Detector with Adaptive Prior and Multi-modal Supervision

Yunxiang Liu u7191378
Shiqiao Zhou u7155524
Kai Xi u6458119

Abstract

Human Object Interaction (HOI) Detection is a task to locate a pair of interacting human and object and classify the interaction type in an image. It is critical to understanding the visual scene semantically. Current Transformer-based end-to-end HOI detectors applies a set of learned embeddings as queries for decoding, which remain constant during inference and do not contain much prior information at the sample level. We propose a novel HOI Grouping layer module to group the feature tokens into an HOI query or instance efficiently. Based on it, we propose two new architectures, utilising the module as a HOI query generator or as a decoder directly. In addition, we fuse the pre-trained verb vectors for actions in the dataset as linguistic query to enhance the prediction of HOI actions. Furthermore, we come up with a new two-dimension (2-D) positional encoding based on 2-D Fourier series for improving the prior location encoding of HOI instances. We conducted experiments to show that the proposed methods may improve the model performance under several different combinations.

1. Introduction

Human Object Interaction (HOI) detection is task aiming to predict a set of $< human, object, interaction >$ tuples from an image. Specifically, given an image input, the detector should output a set of HOI instances, each containing a human box, an object box and its class, and the class of their interaction. In general, the models introduced for this task have evolved from two-stage architectures, to one-stage architectures and then end-to-end architectures which mostly are Transformer-based. Current Transformer-based models are typically encoder-decoder models and usually perform relatively well. Such an architecture requires inputting query vectors to the decoder. However, fusing useful prior knowledge into these vectors are often challenging but critical to the performance. Most of the model proposed so far set these query vectors as learnable embeddings, trained on the dataset level. These embeddings lack instance-level knowledge and their usefulness is likely to be limited during inference.

In this report, we first propose a novel HOI Grouping layer module, which can progressively merge feature representation vectors into groups of arbitrary shapes at any spatial location. It is flexible and compatible with most current Transformer-based detector models. We then propose two new HOI Grouping Network architectures based on it. The first is to apply the HOI Grouping layer as a query vector generator, so that the generated tokens are fused with instance-level details dynamically. For the second architecture, we propose applying the module as a decoder to substitute the original Transformer decoder. This views the HOI detection task as learning a good method to group interacting human and object features to form an HOI representation. It is more flexible than the original Transformer-based models and has great potential in information fusion and model simplification.

Furthermore, we propose fusing pre-trained verb vectors for interaction actions into queries as prior knowledge. Such knowledge aims to enhance interaction classification. We also introduce a new 2-D Fourier series positional encoding in order to improve the localization of the human and the object. Our experiments show that these methods can improve the prediction performance in combination with our baseline model and the two new HOI Grouping Networks.

Our contributions in this report are as follows:

- We propose a novel hierarchical module - HOI Grouping Layer, which can combine connected human and object features into arbitrary groups. We utilise this module and introduce two HOI Grouping Networks: Group-based Query Block as a query vector generator and Group-based Decoder as a new type of decoder.
- We introduce methods to pre-train verb embeddings associating with interaction actions' information from the dataset, and the way to fuse them into queries as prior knowledge.
- We propose a novel positional embedding method based on 2-D Fourier series which is more mathematically supported in theory. This method encapsulates relative relationships of the two dimensions and is more distinguishable compared with the current concatenation embedding.
- Conducting thorough experiments, we show that various aspects of the task performance are improved under

different combinations of our proposed methods. Furthermore, we show that our Group-based decoder architecture may achieve competitive accuracy with fewer parameters and a simper structure.

2. Related Work

2.1. HOI detection

Two-stage and One-stage HOI Detection Modern HOI detection started to get popular with two-stage approaches [2, 7, 10, 14], where the task is considered as a downstream task of object detection. Usually, there is a first stage for object detection and a second stage for multi-steam feature extraction and fusing. Unfortunately, capturing and infusing interaction information is challenging for these models. The dense connection between human and object instances detected in the first stage pose challenges on task complexity [10], and the sequential nature of the two stages limits the efficiency. In contrast, one-stage modules do not need to detect objects explicitly in a separate process. They often utilise creative HOI representations so that the inference can be done in parallel. Such models often have two branches, one for implicit object detection and one for HOI matching prediction. PPDM [11] represents an HOI instance as three points. It defines human and object points as the centers of the bounding boxes, and the interaction point as the mid-point of matching human and object points. Similarly, UnionDet [9] represents HOI instances as the union region of human and object boxes and detects HOI on a union-level in one branch. In the other branch, it detects human and object instances and their localization in the union region. One-stage models are often simpler and more efficient than two-stage models. However, they still often require complicated post processing for fusing the outcomes from different branches.

End-to-End HOI Detection End-to-end models for object detection has been developed over many years. Stewart et al. [16] proposed an LSTM-based method which adopts an encoder-decoder architecture to detect people in crowd. In recent years, DETR [1] improved the architecture by replacing RNN used in previous models with Transformer. Such changes are also reflected in models for HOI detection. QPIC [17] replaced the object detection head of the DETR [1] model with an HOI detection head, introducing a Transformer-based end-to-end HOI detection method. By leveraging the global attention mechanism, Transformer-based HOI detectors can extract and embed humans' and objects' semantic features, their spatial relationship and interaction features all together in one pass. Therefore, all elements of HOI instances can be directly predicted from the output. Such models are even more efficient and simpler than one-stage models, and they can often achieve better accuracy. It has become a new trend of the HOI detection task.

We use HOITrans proposed by Zou et al. [23] as our baseline model. It has a typical Transformer-based HOI detection architecture, as shown in Figure Fig. 1. An image goes through a CNN backbone, gets flattened into vectors, passes through a Transformer encoder and generates the feature representations. Then the feature tokens are queried by a set of learnable embedding - HOI queries in a Transformer decoder. Finally, the output embeddings are passed through an HOI detection head consisted of multiple MLPs, to predict the HOI instances. We choose ResNet-50 as the backbone in our baseline model.

Though it often performs relatively well, the HOI queries passed into the Transformer decoder in this architecture are problematic. Firstly, when trained and tested on HICO-DET [3] dataset, the number of HOI queries is often set to 100, like in [4, 8, 23]. This number is chosen empirically, where in HICO-DET dataset, it is likely that there are at most 100 HOI instances in each image. Such setting is dataset sensitive and its generalisation is limited. Secondly, the HOI queries are learned on the dataset level and remain constant once trained. They do not vary with different input images and therefore cannot reflect input sample's individual context during testing. In this project, we mainly target the second problem by enhancing the query vectors in multiples ways.

2.2. Group Vision Transformer

Features extracted by most computer vision models often remain in a local rectangular region. Their application is limited in tasks requiring more flexible settings. Group Vision Transformer (GroupViT) [20] introduced a hierarchical architecture for the segmentation task, that can partition the feature vectors into groups of any shape progressively. Therefore, connecting pixels can be linked together even if they are not in the same grid. Combined with text supervision, GroupViT achieved state-of-the-art performance on segmentation. We observe similarities between our HOI detection task and the segmentation task. If we view the two bounding boxes in one HOI instance as a whole, they can form any arbitrary shapes with straight edges. Their locations in the image are not limited either. The major difference between the two tasks is that segmentation requires a hard label for each pixel, where our HOI detection task allows multiple-to-multiple relationships. We consider that the global attention mechanism [5] ViT and GroupViT relies on also helps with the chanllenges in the HOI detection task. We modify the Grouping Block in GroupViT to introduce our novel HOI Grouping Layer module. It retains the advantages of flexible shapes and spatial locations, while also adapts to the additional HOI detection task requirement.

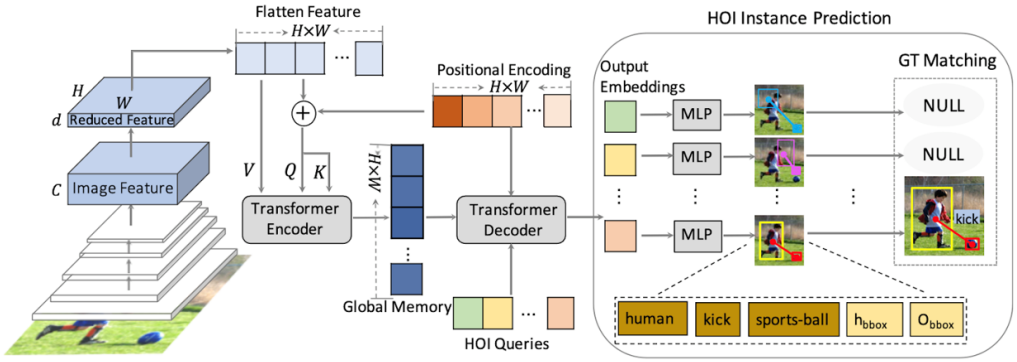


Figure 1. The structure of HOITrans [23]

2.3. Multi-modal Approaches in HOI detection

Language semantics has been widely studied in the field of HOI detection. There are many multi-modal approaches. For example, previous studies have modeled the basic regularity and general relationship of verbs and objects in HOI to solve the challenge of long-tail HOI categories [19]. The authors constructed a knowledge map based on training datasets and real annotations from external sources. Peyre et al. [14] found that detection of visual relation can rely on not only visual phrase information, but also visual-language embedding for subject, object and predicate.

In addition, the Transformer-based methods on HOI detection thrived in recent years. The Transformer decoder in our baseline model transforms three inputs of global image features, positional encodings and HOI queries into output embeddings. However, the HOI query inputs are randomly initialized. Thus, they do not contain much meaningful information at this stage. Inspired by modification on embedding in prior research [22], we introduce a verb fusion block with two inputs: HOI queries and pre-trained training dataset verb embeddings. These verb embeddings are injected with the prior of our training dataset: HICO-DET. Our verb fusion of multi-modal approach makes the HOI queries fit better into the training dataset.

2.4. Positional Embeddings

Transformer, unlike CNN, requires position embedding to encode the location information of tokens, mainly because self-attention is order-invariant. If the location information of token is not provided to the model, the model needs to learn the position through the semantics of tokens, which increases the learning cost. The current two-dimension positional embedding in Transformer model [6, 18, 18] is shown in Eq. (1) and Eq. (2) below. However, this concatenation of two cosine embedding does not show any mathematic support or information exchange between different dimensions. To solve this problem, we coupled the two cosine positional embedding based on the 2-D fourier series, which is discussed in Section 3.3.

$$\begin{aligned} S(x) &= \sum_{n=0}^{d} exp(2\pi f j n x) \\ &= \sum_{n=0}^{d} cos(2\pi f n x) + j sin(2\pi f n x). \end{aligned} \tag{1}$$

$$Pos(x) = \left[cos(\frac{x}{10000^{\frac{n}{d}}}), sin(\frac{x}{10000^{\frac{n}{d}}}) \right]_{n=0:d} \tag{2}$$

3. Methodology

3.1. HOI Grouping Network

We propose a new Group-based mechanism at the decoding stage of the model for the HOI detection task. Figure 2 shows the hierarchical architecture of the HOI Grouping layer module applying this mechanism. It takes the feature representation tokens and a set of group embeddings, and then gradually combine the features into arbitrary-shaped groups. A Transformer encoder is added before each HOI Grouping layer to exploit the global self-attention for better group representations. The HOI Grouping layer module is flexible and can be applied in combination with current Transformer-based encoder-decoder architecture in multiple ways. We mainly apply the HOI Grouping layer module in two ways. We first try to use it as an enhancement to the HOI queries inputted into the Transformer decoder. In this application, the Group-based mechanism is used as a method to generate HOI queries containing image sample's individual context dynamically. We name it the Group-based Query Block and discuss it in detail later. Then we try to replace the whole Transformer decoder with our new HOI Grouping layer module. In this case, it utilizes the Group-based mechanism to generate good potential HOI instance embeddings directly. It is named as Group-based Decoder.

HOI Grouping Layer As shown in figure Fig. 2 , the HOI Grouping Layer takes group embeddings and feature representation tokens as input and outputs the grouped features

after two stages. In the first stage, it enhances the group embeddings with a 6-head global attention, using the original group embeddings as queries and the feature representations as keys and values. At the second stage, it merges the original feature tokens that are assigned to the same group into a single new grouped feature token, according to the similarities to the enhanced group embeddingss.

Formally, we compute the enhanced group embeddings EG in the first stage as

$$EG = G + \text{6-Head-Attention}(Q_1, K_1, V_1) \quad (3)$$

$$Q_1 = GW_{q1}, K_1 = FW_{k1}, V_1 = FW_{v1}$$

where G, F are original group embeddings and feature representations respectively.

In the second stage, we combine the original feature tokens into grouped features GF with

$$GF = EG + \text{1-Head-Attention}(Q_2, K_2, V_2) \quad (4)$$

$$Q_2 = EG \cdot W_{q2}, K = FW_{k2}, V = FW_{v2}$$

Specifically, each feature token can be added to multiple groups non-exclusively in the second stage. This is because for HOI detection task, one person may interact with multiple object and one object may also interact with multiple humans. Instead of assigning a hard group label for each feature token as segmentation Grouping Block in [20] does, we merge the features by softmax scores across keys. Formally, the attention score matrix $A_{i,j}$ between the (enhanced) group embedding g_i in the second stage and the original feature token f_j is computed as

$$A_{i,j} = \frac{exp(W_{q2}g_i \cdot W_{k2}f_j)}{\sum_{k=1}^{M} exp(W_{q2}g_i \cdot W_{k2}f_k)} \quad (5)$$

where M is the total number of the original feature tokens and W_{q2} and W_{k2} are the weights of the learned linear projections for attention in the second stage. Note that $\sum_{j=1}^{M} A_{i,j} = 1$ and $0 \leq \sum_{i=1}^{N} A_{i,j} \leq N$, where N is the total number of groups.

Group-based Query Block The first way to add our new Group-based mechanism to the baseline model is to use it as a method to propose HOI queries. This is called the Group-based Query Block in figure Fig. 3a . This block fuses image features with two hierarchical grouping queries, first 256 and then 100, to create 100 HOI queries. In this case, the HOI queries are not constant anymore; They capture the individual features of the sample image with the Group-based mechanism and therefore should be more dynamic and general than embeddings learned at the database-level. This specifically targets and solves the second problem of HOI queries we discussed in Section 2.1 "End-to-End HOI Detection". Note that we keep the empirical setting of 100 HOI queries in total in this project.

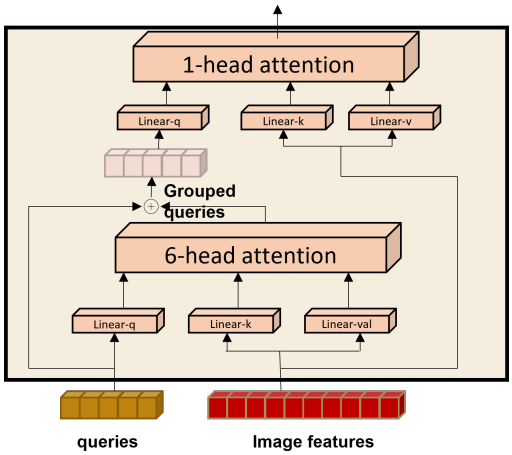
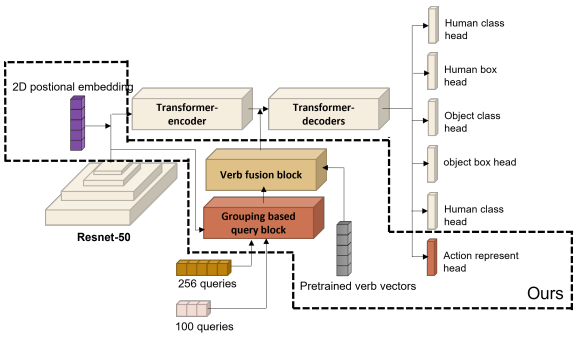


Figure 2. The HOI Grouping Layer structure

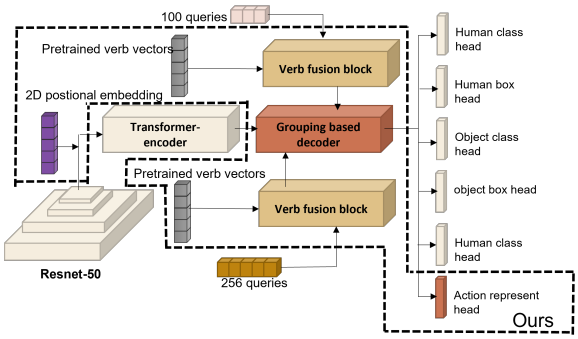
Group-based Decoder Another way to apply the new Group-based mechanism is using it as a new decoder to replace the original Transformer decoder in the baseline model. This is as shown in the figure Fig. 3b, where we group the original features into 256 groups, refine with 256 grouping once and then group them into 100 output embeddings. In this case, the HOI task can be viewed as a task to find a good way to group feature tokens into potential HOI instances, with arbitrary shapes at any spatial locations. This structure is just built upon the new Group-based mechanism, without repeatedly querying the original feature tokens at each decoder layer. Compared to the Transformer decoder, our new Group-based decoder is a lot more flexible. As there is no limit on how many groups to be used in each HOI Grouping layer, the model can either be simplified with more coarse grouping or be enhanced with finer grouping at each layer. With our current setting (256, 256, 100), we attain a simpler model with less parameters than the baseline model. Furthermore, each HOI Grouping layer accepts a new set of group embedding inputs. Compared to the Transformer decoder where group embeddings can only be fused at the very bottom layer, we can fuse prior knowledge into our new decoder at different levels, which further increases the flexibility and the potential of the model.

3.2. Multi-Modal Query

Current pre-trained word embeddings come from pre-trained model *e.g.* GloVe [13], which contains the co-occurrence priors from *e.g.* Google News. The information from these word embeddings may not fit in our HICO-DET dataset. Insipired by Verb Semantic Model (VSM) [22], we use VSM by verb semantic reasoning and SKL Loss to inject HICO-DET co-occurrence priors. Then we fuse the new word embedding with queries at the transformer decoding stage instead of the predicting stage.



(a) structure of Group-based Query Block



(b) structure of Group-based Decoder.

Figure 3. Our two proposed structures based on the basline, the region surrounded by dotted line are our modifications

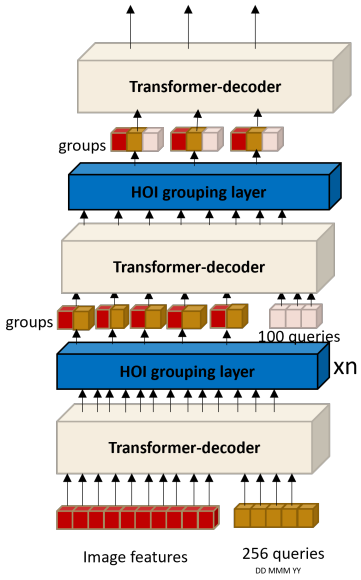


Figure 4. the detailed structure of Grouping based query block and group based decoder, n equals to 2 for Grouping based query block and equals to 1 for group based decoder

Pre-trained Verb Vector To inject the training dataset prior into word embeddings, we apply Verb Semantic Model (VSM) to retrain the pre-trained word embeddings. The first step of VSM is verb semantic reasoning. Here we adopt a transformation. Suppose the original word embedding is $\mathbf{P} = \{\mathbf{p}_i \in \mathbb{R}^{D_p}\}_{i=1}^{N_p}$, where D_p and N_p denote the number of verbs in the HOI dataset and the dimension of word embeddings. We gain $\widetilde{\mathbf{P}} = \{\widetilde{\mathbf{p}}_i \in \mathbb{R}^{D_p}\}_{i=1}^{N_p}$ by verb semantic reasoning, as shown in equations Eq. (6) and Eq. (7). In Eq. (6), r_{ij} represents softmax function with index j on the pairwise relation between $\mathbf{p}_i$ and $\mathbf{p}_j$; θ and ϕ represent linear projection functions; In Eq. (7), σ represents ReLU; $\mathbf{W} \in \mathbb{R}^{DxD_p}$ represents the linear projections of embedding and residual connection.

$$r_{ij} = softmax_j(\frac{\theta(p_i)^T \phi(p_j)}{\sqrt{2}}) \tag{6}$$

$$\widetilde{\mathbf{p}}_i = \sigma\left(\sum_{j=1}^{N_p} r_{ij}\mathbf{W}_{p1}\mathbf{p}_j\right) + \mathbf{W}_{p2}\mathbf{p}_i \tag{7}$$

The second step is to apply SKL loss to push adjacency matrix of $\widetilde{\mathbf{P}}$ to co-occurrence distribution of HICO-DET dataset. Here we need to extract conditional probabilities of verbs in the HICO-DET dataset. From prior research [21], we denoted all verbs in HICO-DET dataset as $V = \{i\}_{i=1}^{N_p}$. Thus we defined C as conditional probabilities set in HICO-DET dataset. This is shown in Eq. (8). The symmetrized conditional probability distribution $\hat{C}$ is defined in Eq. (9), where $2N_p$ is used for normalization. Then we gain the adjacency matrix A in Eq. (10) from $\widetilde{\mathbf{P}}$, where τ is temperature parameter to adjust the softmax value.

$$C = \{c_{ij} = p(j|i)|i, j \in V, i \neq j\} \tag{8}$$

$$\hat{c}_{ij} = \frac{c_{ij}the + c_{ji}}{2N_p} \tag{9}$$

$$A = \{a_{ij} = \frac{exp(\widetilde{p}_i^T \widetilde{p}_j/\tau)}{\sum_{k=1}^{N_p}\sum_{l=1,l\neq k}^{N_p} exp(\widetilde{p}_k^T \widetilde{p}_l/\tau)}|i, j \in V, i \neq j\} \tag{10}$$

Finally, we leverage KL-divergence to push the distribution A to the distribution $\hat{C}$. Eq. (11) is the loss function from KL-divergence. We apply this loss function to back-propagation to supervise the retraining of our pre-trained embedding, so that the co-occurrence distribution of HICO-DET dataset is injected into embeddings.

$$L_{SKL} = \mathbb{E}_{\hat{C}}[\log(\hat{C}) - \log(A)] \tag{11}$$

Linguistic queries fusion with multi modal supervision Here we introduce a new verb fusion block. After we gain

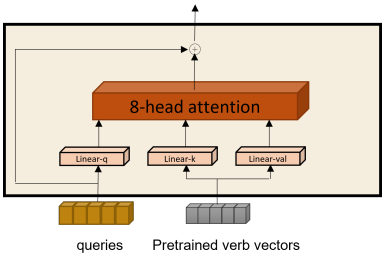


Figure 5. structure of linguistic fuson block

the new pre-trained embeddings, we fuse these language features into our model. Different from prior HOI detectors, we fuse the language features with original HOI queries in the transformer decoder stage instead of prediction stage.

The structure of our verb fusion block is shown as below Fig. 5. To supervise the verb fusion, we add contrastive loss with four loss parts: matched verbs with all queries, matched queries with all verbs, matched verbs with matched queries and matched queries with matched verbs. These four parts supervise verb fusion in different scales. The Eq. (12) shows the contrastive loss formula inspired by attention mechanism, where $i, j \in M, i \neq j$; N is number of queries and M is matched number between queries and verbs; τ is temperature parameter to scale the value; q and k are the query and the key in the attention mechanism.

$$Contrastive(q_N, k_M) = -\frac{1}{N}\sum^{N}\log\frac{exp(\tau q_n^T k_n)}{\sum_M exp(q_i^T k_j)} \tag{12}$$

3.3. Two-dimension Positional Encoding

The current strategy of two dimension positional embedding is a concatenation of two one-dimension embedding based on fourier series from the formula Eq. (1). However, the current method is a simple collection of two 1-D positional representation. In fact, as an single point presents in a 2-D space, the x and y coordinates should be represented as a whole. As a result, we rethink how the Eq. (2) is derived from Eq. (1) and get a new positional vector from the two-dimension version of Eq. (1), as Eq. (13) shows:

$$S(x, y) = Pos(x)Pos(y)^{\mathbf{T}} \tag{13}$$

From Eq. (13), we could verify that for each element in postion x, it will interact with every element in position y. Although not strict mathmatically, it is more implementable and is approaching the interaction between the two dimensions. The detailed derivation of our proposed method will be shown in Appendix A.

Algorithm 1: Pseudo code for 2-D positional embedding

Input: input parameters Position x, Position y
Output: output result, 2-D positional embedding $\mathbf{v} \in \mathbb{R}^{d^2}$
Symbols: 1-D position embedding function $p(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}^d$

$\mathbf{v}_x = p(x)$;
$\mathbf{v}_y = p(y)$;
resize $\mathbf{v}_x$: $\mathbb{R}^d \longrightarrow \mathbb{R}^{d\times 1}$;
resize $\mathbf{v}_y$: $\mathbb{R}^d \longrightarrow \mathbb{R}^{1\times d}$;
element-wise multiplication: $\mathbf{v} = \mathbf{v}_x \odot \mathbf{v}_y \in \mathbb{R}^{d\times d}$;
flatten $\mathbf{v} : \mathbb{R}^{d\times d} \longrightarrow \mathbb{R}^{d^2}$;
return $\mathbf{v}$;

4. Experiments

To verify the effectiveness of our model, we experiment with our proposed methods on HICO benchmarks. The following subsections introduce the dataset, discuss the experiment details, compare our approaches among all submodules with other published methods and do ablation studies to evaluate the act of each module.

4.1. Datasets

We train and evaluate all of our proposed method on HICO-DET [3] dataset, which contains 47,774 images (9,658 samples among those for testing). The dataset contains 80 object classes and 117 action classes from MS-COCO [12], including an additional "no interaction" class. In addition, if the number of images of an action is less than 10 in the dataset, it will be regarded as a hard example.

4.2. Metrics

We evaluate our model performance mainly by mAP scores, which is the mean average precision among all action classes. For calculating mAP, we first calculate the points on Precision-Recall (PR) curve for all classes. Then we split the curve into 10 parts, and compute the mAP by mean of all classes.

4.3. Implementation Detail

During training, we set most of the hyper-parameters the same as our baseline model [23]. Since the models contain relatively large Transformer structures, we train them on 4x NVIDIA RTX 3090 GPUs. To reduce the time and the number of epochs required, we download the backbone weight from the baseline and freeze it during training. We implement the AdamW optimizer with the starting learning rate as 1e-4 and the weight decay as 1e-4, where the learning rate is dropped to 5e-5 after 110 epochs. We train each

of our proposed models for 150 epochs, and the total training time required for each model is roughly 41 hours. For reproduction of baseline model, please check **??** for detail.

4.4. Comparison with current methods

We mainly report quantitative results from action recognition papers recommended by the lecturer and some current research projects on HICO-DET dataset, as shown in Tab. 1. Our method using Group-based query block achieve 36.29% improvement over the best two-stage method on Full categories, 24.68% improvement on Rare categories and 38.14% improvement on NonRare categories. Compared with one-stage and multi-modal methods, our two methods have also improved significantly.

For end-to-end methods, we firstly reproduce the results of the baseline model. To be fair, the baseline model retrained by us achieves a better performance than the results in the original paper [23]. The reason may be that our hardware training environment is better than what original paper used. Our Group-based Query Block model improves the performance on three categories by around 10% over the baseline model (paper). Specifically, our method's performance has an improvement in Rare categories, but a slight decrease in NonRare categories.

4.5. Comparison study

We compare all our proposed methods and the baseline to verify the act of each modification. The results are shown in Tab. 2.

Group-based Query Block This block enhances the model detection on rare HOI instances. The results in Tab. 2 indicate that the block is robust at combining regions of interest (ROI) which belong to rare interaction of people and objects. However, for the NonRare samples detection, it shows a worse performance than the baseline. The reason may be that the dynamic encoding of individual images reduces the model ability of representation on the overall dataset.

Group-based Decoder Group-based decoder has less number of parameters due to the model architecture. However, all the metrics in Tab. 2 have performance reduction compared with baseline. This is not only due to the less number of parameters but also the large number of queries, which increases the difficulty of the model learning for prior. Besides, the Group-based Decoder still acts more like a encoder. Therefore, switching the transformer decoder with Group-based Decoder may lower the model performance on information decoding.

Multi-modal query After implementing the multi-modal approach on baseline, we find that the model inference performance on rare samplers increase. However, the Tab. 2 also shows that this modification is not good at testing normal examples, which leads to the drop of the overall mAP. One reason is that the NonRare object is in the majority of the HICO-DET dataset. Another reason may be that introduction of linguistic modal enriches the feature representation related to actions but may make the model overfit easy examples. In contrast, for rare actions, the model does not learn strong features due to the small number of rare actions. Thus, introducing the richer word representation makes the model fit well on these samples.

Method	Backbone	mAP		
		Full	Rare	NonRare
Two-stage methods				
HO-RCNN [2]	CaffeNet	7.81	5.37	8.54
InteractNet [7]	ResNet-50-FPN	9.94	7.16	10.77
TIN [10]	ResNet-50	17.22	13.51	18.32
Peyre et al. [15]	ResNet-50-FPN	19.40	14.63	20.87
One-stage methods				
PPDM [11]	Hourglass-104	21.73	13.78	24.10
UnionDet [9]	ResNet50-FPN	17.58	11.72	19.33
Multi-modal methods				
Xu et al. [19]	ResNet-50	14.70	13.26	15.13
End-to-end methods				
Baseline (Paper) [23]	ResNet-50	23.46	16.91	25.41
Baseline (Retrained by Us)	ResNet-50	26.44	17.85	29.01
Baseline + Group-based Query Block (**Ours**)	ResNet-50	26.44	18.24	28.83
Baseline + Group-based Decoder (**Ours**)	ResNet-50	25.57	16.44	28.29

Table 1. Comparison with the two-stage, one-stage, multi-modal and end-to-end methods on HICO-DET test set [3].

methods	param(M)	mAP			Recall
		Full	Rare	NonRare	
Baseline	41.46	23.46	16.91	25.41	-
Baseline (retrained by us)	41.46	26.44	17.85	29.01	52.95
Baseline+Group-based Block	48.92	26.44	18.24	28.83	51.79
Baseline+Group-based Decoder	40.77	25.57	16.44	28.29	47.87
Baseline+multi-modal Query	42.32	26.43	18.24	28.88	51.79
Baseline + 2-D Positional Embedding	41.46	26.49	17.53	29.17	53.11

Table 2. The comparison among our proposed work with baseline on HICO-DET test set [3].

Multi-Modal Query	2-D Positional Embedding	mAP			Recall
		Full	Rare	NonRare	
		25.57	16.44	28.29	47.87
✓		24.13	16.94	26.44	50.52
	✓	26.27	17.29	28.61	48.32
✓	✓	25.13	16.86	27.61	48.89

Table 3. Ablation experiments for our model with Group-based decoder. Implementation details are shown in Section 4.3.

4.6. Ablation Study

In our ablation study, we focus on the simpler Group-based Decoder models due to the resource limitation. We verify the performance of both the language fusion block and the 2-D positional embedding.

Effectiveness of Multi-modal supervision To improve the robustness of pre-trained word vectors, we normalize it before input them into model. The number of verb vectors is 117, which is same as HICO-DET dataset. As the Fig. 3b shows, we fuse verb vectors with two sets of queries produced by the standard Embedding layer. And the initial

temperature of contrastive loss is set to be one. After testing, we find that the language module improves the mAP score on rare samples and the Recall metric, which is similar to the comparison study results. The multi-modal query can indeed improve the mAP score in rare samples and the Recall by successful prediction on rarer samples.

Importance of two-dimension positional embedding According to the encoding mechanism showed in Section 2.4. The dimension of each 1-d positional embedding is set to 16. Therefore, the dimension of resulted positional embedding will be 256. The results in Tab. 3 show that positional embedding improves the overall mAPs of model in all types of HOI instances but reduces the Recall. The reason may be that stronger positional representation improves the precision of object location but fails to find some specific types of objects.

5. Discussions

Figure 6. Our *correct* results on some HICO-DET test images.

Figure 7. Our *false* results on some HICO-DET test images.

Qualitative Results We show the detection results from our Group-based Decoder model in Figure 6. Our model can overcome many common challenges in the HOI detection task and successfully detect HOI instances under some hard scenarios. One of the common challenges in the task is multiple interactions between one or more humans and one or more objects, through one or multiple different actions. Image A, B and C show that our model can distinguish multiple interactions in the same image well, even if they are spatially close or semantically confusing. Image D shows that our model can still successfully infer the position of the human from the context even if the human is partially or fully blocked. Image E shows that our model can also detect the interaction between a human and an object far away. Furthermore, our model can also successfully detect the lack of interaction in the input, like in image F.

Failure Cases Figure 7 shows some failure cases produced by our model. We select 6 representative images including both false positive cases (images G, H, I) and false negative cases (images J, K, L). Our model may confuse the object class (*e.g.* image G), mistake a human-like object as a human (*e.g.* image H) or mis-classify the interaction between two humans (*e.g.* image I). In the false negative cases, our model has difficulty detecting the object when it overlaps with the human and/or gets blocked (*e.g.* image J, K). Sometimes, it may also have trouble detecting HOI instances in dimly colored images (*e.g.* image L).

The overlapping between the human and the object in the same HOI instance may be particularly challenging to our Group-based Decoder model. In the group-based mechanism, each token can be added to each group for at most once. As a result, the feature token in the overlapped region cannot be emphasized more than the tokens in the non-overlapping areas. Therefore, such tokens may be less distinguished compared to the Transformer decoder, where the feature tokens can be queried multiple times and emphasized to varying degrees.

6. Conclusion

In conclusion, we proposed a novel hierarchical module - HOI Grouping Layer, and develop two model architectures, Group-based Query Block and Group-based Decoder based on that. With the assistance of multi-modal fusion strategy and positional encoding, the model has improvements on different aspects compare to the baseline model.

Although our method has some achievements as discussed in Section 5, the Group-based Decoder is not good at dealing with HOI instances with high overlapping. Thus, for future work, we decide to do more investigation on our first model architecture (Group-based Query Block). Besides, for introducing multi-modal information on training stage, we need to use more of model predictions of word embedding branch during inference. Finally, we will develop the box representation based on 4-D fourier series to present the box priors, which will give possibilities to introduce the anchor boxes into Transformer-based detectors.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *workshop on applications of computer vision*, 2017.

[3] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[4] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[7] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. *computer vision and pattern recognition*, 2017.

[8] 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.

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

[10] 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. *computer vision and pattern recognition*, 2019.

[11] 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.

[12] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[14] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. *international conference on computer vision*, 2018.

[15] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1981–1990, 2019.

[16] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016.

[17] 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.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[20] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.

[21] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12709–12716, 2020.

[22] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics, 2022.

[23] 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.

Appendices

A. Derivation of 2-D position embedding based on 2-D fourier series

Let's recall how current 1-D postional embedding get based on fourier series.

Let's have the 1-D fourier series as:

$$\begin{aligned} S(x) &= \sum_{n=0}^{d} exp(2\pi f j n x) \\ &= \sum_{n=0}^{d} cos(2\pi f n x) + j sin(2\pi f n x). \end{aligned} \tag{14}$$

Then, we switch the cumulative symbol into a vector, then we get Eq. (15):

$$\begin{aligned} Pos(x) = &[cos(2\pi f x), sin(2\pi f x), cos(2\pi 2 f x), sin(2\pi 2 f x), \\ &\cdots, cos(2\pi n f x), sin(2\pi n f x)]^{\mathbf{T}} \end{aligned} \tag{15}$$

from Eq. (15), we could see that every rank in Eq. (14) will be the every elment in Eq. (15)

We could follow this process to devrite the 2-D positional embedding which follows:

$$\begin{aligned} S(x,y) &= \sum_{m=1}^{d_1}\sum_{n=1}^{d_2} exp(2\pi f j (mx + ny)) \\ &= \sum_{m=1}^{d_1}\sum_{n=1}^{d_2} exp(2\pi f j m x) exp(2\pi f j n y) \end{aligned} \tag{16}$$

combine Eq. (14) and Eq. (16), we have:

$$\begin{aligned} S(x,y) = \sum_{m=1}^{d_1}\sum_{n=1}^{d_2} &cos(2\pi f m x) cos(2\pi f n y) \\ &+ j cos(2\pi f m x) sin(2\pi f n y) \\ &+ j sin(2\pi f m x) cos(2\pi f n y) \\ &- sin(2\pi f m x) sin(2\pi f n y) \end{aligned} \tag{17}$$

Follows the same process as 1-D positional embedding, we can devrite the Eq. (17) as a 4-D tensor as:

$$Pos(x,y) = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m,1} & p_{n,2} & \cdots & p_{m,n} \end{pmatrix} \tag{18}$$

for each element in Eq. (18), we have:

$$p_{i,j} = \begin{pmatrix} cos(2fix)cos(2fjy) & cos(2fix)sin(2fjy) \\ sin(2fix)cos(2fjy) & -sin(2fix)sin(2fjy) \end{pmatrix} \tag{19}$$

We then could flatten the tensor to get final 1-D representation. However, in practise, get the vectors shown on from Eq. (18) and Eq. (19) is hard for implement, for simplfy, we do a engineering approaching through combining two 1-D positional embedding shown in Eq. (15) as:

$$\begin{aligned} Pos(x,y)^{'} &= Pos(x) Pos(y)^{\mathbf{T}} \\ &= \begin{pmatrix} p^{'}_{1,1} & p^{'}_{1,2} & \cdots & p^{'}_{1,n} \\ p^{'}_{2,1} & p^{'}_{2,2} & \cdots & p^{'}_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ p^{'}_{m,1} & p^{'}_{n,2} & \cdots & p^{'}_{m,n} \end{pmatrix} \end{aligned} \tag{20}$$

for each element in Eq. (20), for $i \in [1, 1 + m/2], j \in [1, 1 + n/2]$, we have:

$$p^{'}_{2i-1,2j-1} = cos(2\pi f(2i-1)x) cos(2\pi f(2j-1)y) \tag{21}$$

$$p^{'}_{2i-1,2j} = cos(2\pi f(2i-1)x) sin(2\pi f 2jy) \tag{22}$$

$$p^{'}_{2i,2j-1} = sin(2\pi f 2ix) cos(2\pi f(2j-1)y) \tag{23}$$

$$p^{'}_{2i,2j} = sin(2\pi f 2ix) sin(2\pi f 2jy) \tag{24}$$