

Hierarchical Memory Learning for Fine-Grained Scene Graph Generation

Youming Deng¹ Yansheng Li¹(✉) Yongjun Zhang¹ Xiang Xiang²
Jian Wang³ Jingdong Chen³ Jiayi Ma⁴

¹School of Remote Sensing and Information Engineering, Wuhan University

²School of Artificial Intelligence and Automation, Huazhong University of Science and Technology ³Ant Group

⁴Electronic Information School, Wuhan University

Abstract. Regarding Scene Graph Generation (SGG), coarse and fine predicates mix in the dataset due to the crowd-sourced labeling, and the long-tail problem is also pronounced. Given this tricky situation, many existing SGG methods treat the predicates equally and learn the model under the supervision of mixed-granularity predicates in one stage, leading to relatively coarse predictions. In order to alleviate the impact of the suboptimum mixed-granularity annotation and long-tail effect problems, this paper proposes a novel Hierarchical Memory Learning (HML) framework to learn the model from simple to complex, which is similar to the human beings' hierarchical memory learning process. After the autonomous partition of coarse and fine predicates, the model is first trained on the coarse predicates and then learns the fine predicates. In order to realize this hierarchical learning pattern, this paper, for the first time, formulates the HML framework using the new Concept Reconstruction (CR) and Model Reconstruction (MR) constraints. It is worth noticing that the HML framework can be taken as one general optimization strategy to improve various SGG models, and significant improvement can be achieved on the SGG benchmark.

Keywords: Scene Graph Generation, Mixed-Granularity Annotation, Hierarchical Memory Learning

1 Introduction

The task of Scene Graph Generation (SGG) [51] is a combination of visual object detection and relationship (i.e., predicate) recognition between visual objects. It builds up the bridge between computer vision and natural language. SGG receives increasing attention since an ideal informative scene graph has a huge potential for various downstream tasks such as image caption [17,54] and VQA [1,31]. To pursue the practical application value, SGG models keep working towards generating an informative scene graph where the fine-grained relationship between two objects should be predicted. Earlier works like [43,53,57] only design models for feature refinement and better representation. However, they

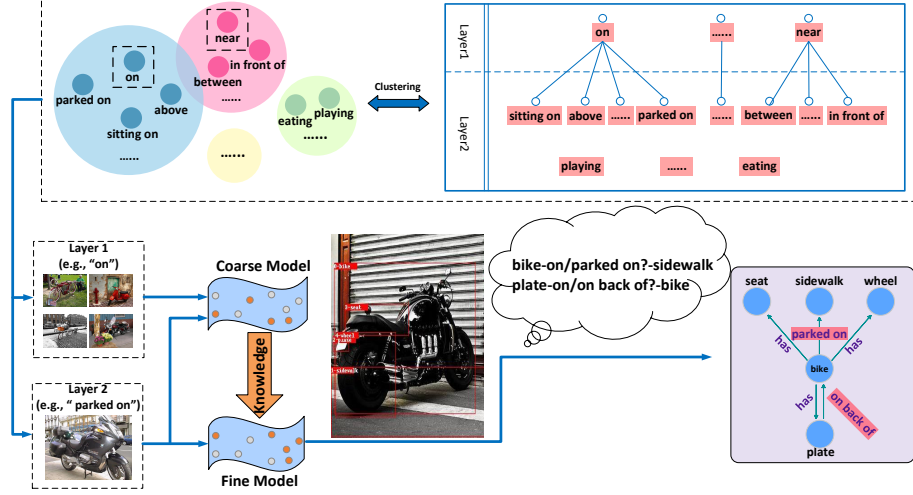


Fig. 1: **Automatic Predicate Tree Construction and visualization of the HML Framework.** After constructing a hierarchical predicate tree via clustering, the model is trained with the HML framework to make the fine-grained predicate prediction.

ignore the dataset and task properties, limiting the performance. In order to deal with the long-tail effect and various biases [42] within the dataset, recent works including [7, 16, 39, 42] move towards designing the learning framework to improve the overall performance of several classic SGF models. Even with this progress, making the fine-grained predicate prediction is still challenging.

Generally speaking, two factors lead to a frustrating result. The first is the mixed-granularity predicates caused by artificial subjective annotation. The predicate recognition in the dataset is much trickier than the image classification tasks. For instance, although Microsoft coco [29] has super categories, it does not require a model to have the ability to make a different prediction like “bus” and “vehicle” for a visually identical object. However, in the SGF task, a more general predicate like “on” and a more informative one like “parked on” will be learned and predicted simultaneously. Under this training condition, it is a dilemma for models since machines cannot understand why almost identical visual features have different annotations and require them to make different prediction results. The second one is the long-tail effect which exists objectively in nature. Some of the dominant predicate classes are almost 1,000 times as many as less-frequent ones, leading to a bad performance on those less frequent predicates. Too many general predicates like “on” in training will lead to insufficient training for less-frequent ones like “eating” and drifting preference away from fine ones such as “parked on”. Some methods, including re-weighting and re-sampling, seem to be suitable choices. However, due to the hierarchical annotation in the dataset, the improvement seems to be limited according to [42].

When training the deep network, problems like the long-tail effect and mixed-granularity annotation are universal but catastrophic to the deep network training. In contrast, humans seem capable of handling these complicated problems. As shown in cognitive psychology research [14], human beings appear to learn gradually and hierarchically. Coincidentally, the semantic structure of predicates in VG is similar to our real-life naming, consisting of mixed-granularity information. For instance, parents often teach their kids to recognize “bird” first and then go for the specific kinds like “mockingbird” or “cuckoo”. Inspired by this, we realize that it is plausible to design a hierarchical training framework to resolve the abovementioned problems by imitating human learning behavior.

This work proposes a Hierarchical Memory Learning (HML) framework for hierarchical fashion training with the abovementioned consideration. At the very beginning, we cluster predicates, establish a hierarchical tree in Fig. 1 and separate the dataset by the tree layers without any extra manual annotation. To realize hierarchical training, Concept Reconstruction (CR) is used to inherit the previous model’s predicate recognition ability by imitating its output. For a similar purpose, Model Reconstruction (MR) directly fits the parameters in the previous model as a stronger constraint. Under this training scenario, the model gets less chance to confront the previously discussed dilemma and is much easier to train with a relatively small and balanced fraction of predicates.

The proposed HML framework is a general training framework and can train any off-the-shelf SGG model. Fig. 1 shows the scene graph generated by the hierarchical training scenario. The scene graph predicted by the model trained with the HML framework is more comprehensive and fine-grained. The predicted relationships such as “bike-parked on-sidewalk” and “plate-on back of-bike” are more informative and meaningful than “bike-on-sidewalk” and “plate-on-bike”.

The main contributions of this work can be summarized as follows:

- Inspired by human learning behavior, we propose a novel HML framework, and its generality can be demonstrated by applying it to various classic models.
- We present two new CR and MR constraints to consolidate knowledge from coarse to fine.
- Our HML framework overperforms all existing optimization frameworks. Besides, one standard model trained under HML will also be competitive among various SGG methods with the trade-off between fine and coarse prediction.

2 Related Work

2.1 Scene Graph Generation

SGG [44,51,57] has received increasing attention in the computer vision community because of its potential in various down-stream visual tasks [19,46,54]. Nevertheless, the prerequisite is the generation of fine-grained and informative scene graphs. Recent works consider SGG mainly from three perspectives.

Model Design. Initially, some works designed elaborate structures for better feature refinement. [51] leveraged GRUs to pass messages between edges and nodes. [59] explored that the feature of objects and predicates can be represented in low-dimensional space, which inspired works like [20,50,36]. [57] chose BiLSTM for object and predicate context encoding. [43] encoded hierarchical and parallel relationships between objects and carried out a scoring matrix to find the existence of relationships between objects. Unfortunately, the improvement is limited to elaborate model design alone.

Framework Formulation. Later works tried to design the optimization framework to improve the model performance further. Based on causal inference, [42] used Total Direct Effect for unbiased SGG. [39] proposed an energy-based constraint to learn predicates in small numbers. [56] formulated the predicate tree structure and used tree-based class-balance loss for training. [56] and our work both focus on the granularity of predicates and share some similarities.

Dataset Property. The long-tail effect was particularly pronounced in VG, making studying this problem very important. [7] utilized dynamic frequency for the better training. [9] proposed a novel class-balance sampling strategy to capture entities and predicates distributions. [16] sought a semantic level balance of predicates. [27] used bipartite GNN and bi-level data re-sampling strategy to alleviate the imbalance. However, another problem (mixed-granularity annotation) in the dataset is not fully explored, which inspires this work. Our concurrent work [11] also borrowed the incremental idea to overcome this problem. We add semantic information for the separation and stronger distill constraint for better head knowledge preserving, while [11] learns better at the tail part.

2.2 Long-Tail Learning

Only a few works like [7,9,16,27] cast importance on the long-tail effect in VG. In fact, many long-tail learning strategies can be used in SGG. The previous works tackling the long-tail effect can be roughly divided into three strategies.

Re-sampling. Re-sampling is one of the most popular methods to resolve class imbalance. Simple methods like random over or under-sampling lead to overfitting the tail and degrading the head. Thus, recent work like [13,21,47,61] monitored optimization process of depending only instance balance.

Cost-sensitive Learning. Cost-sensitive learning realizes class balance by adjusting loss for different classes during training. [8,29,38] leveraged label frequency to adjust loss and prediction during the training. [40] regarded one positive sample as a negative sample for other classes in calculating softmax or sigmoid cross-entropy loss. Other works [4,22] tried to handle the long-tail problem by adjusting distances between representation features and the model classifier for different classes.

Transfer or Incremental Learning. Those methods help to transfer information or knowledge from head to tail and enhance models' performances. [30,55] proposed a guiding feature in the head to augment tail learning. [48] learned to map few-shot model parameters to many-shot ones for better training. Works like [15,28,32] helped to distill knowledge directly from the head to the tail.

3 Approach

We will first introduce how to automatically construct a hierarchical predicate tree via clustering (Sec. 3.1). And then turn back to explain our HML framework in (Sec. 3.2), along with loss formulation for CR (Sec. 3.3) and MR (Sec. 3.4).

3.1 Soft Construction of Predicate Tree

In order to form a predicate tree for our training scenario, we firstly embed all 50 predicates in the dataset into feature vectors with the pre-trained word representation model in [33]. After that, motivated by reporting bias [34], we cluster predicates into different groups and do some soft manual pruning (e.g., re-classifying some mistakes into different groups). We finally pick up the top-K frequent predicates within each group as the first K layers of the tree.

As for clustering, we choose the traditional distributed word embedding (DWE) algorithm [33], since we wish to eliminate the context from objects or subjects which can provide extra information [57] to the predicate embedding. We iterate through all 50 predicates from frequent to less-frequent for clustering. The first predicate is automatically divided into the first group and records its embedding vector to be the initial value of the first group representation $R_1 = DWE(x_1)$. As for the following ones, we calculate the cosine distance among all current group representations:

$$SS^{ij} = \frac{R_i \cdot DWE(x_j)}{\|R_i\| \times \|DWE(x_j)\|}, \quad (1)$$

where SS^{ij} is the semantic similarity between current iterated predicate x_j and i^{th} group representations R_i . $DWE(x)$ represents distributed word embedding function on the predicate. Max cosine distance SS_{max}^{ij} will be recorded and compared with the empirical threshold T_{SS} whose setting is mentioned in Sec. 4.3. If SS^{ij} is larger than the threshold, the currently iterated predicate x_j will be added to the existing group. Otherwise, we create a new group for it. The group representations will be updated in the following rule:

$$\begin{cases} R_{N+1} = DWE(x_j), SS_{max}^{ij} < T_{SS} \\ R_i = \frac{n \cdot R_i + DWE(x_j)}{n+1}, SS_{max}^{ij} \geq T_{SS} \end{cases}, \quad (2)$$

where N is the current number of groups and n is the number of predicates in the i^{th} group.

After clustering, the most frequent predicates are assigned to the first layer, the second frequent predicates are assigned to the second layer, etc. It is worth noticing that during this clustering, some human actions such as “looking at”, “playing”, “says”, “eating”, and “walking in” will become a single group as one single predicate. We automatically divide them into the last layer since those human action predicates are almost 100 to 1000 times less than predicates in the other layer. The most suitable number of layers depends on the dataset itself. Sec. 4.6 analyzes the best layer number for the VG dataset.

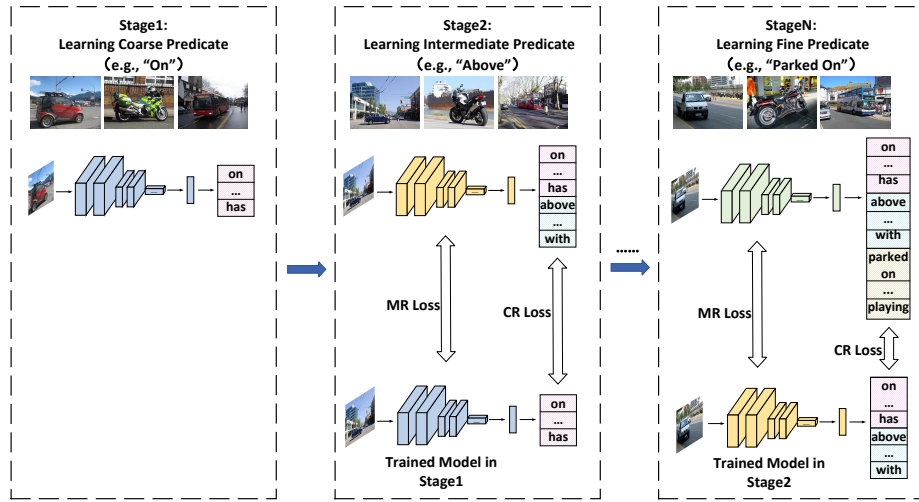


Fig. 2: **Overview of HML Framework.** We train the model in a coarse to fine fashion. In each step, the model calculates CR and MR for knowledge assimilation. Meanwhile, the importance scores and empirical Fisher Information Matrix are calculated after updating the model. The importance score and empirical Fisher Information Matrix are passed down at the end of each stage. For the VG dataset, we set the stage number to be 2 and explain in Sec. 4.6.

3.2 Hierarchical Memory Learning Framework

Most SGG models comprise two steps. In the beginning, an image is fed into an ordinary object detector to get bounding boxes, corresponding features of these regions, and the logits over each object class. These detection results are used to predict the scene graph in the next step. The feature of a node is initialized by box features, object labels, and position. Some structures like LSTM [37,57] are used to refine nodes' features by passing and incorporating the messages. After that, the object labels are obtained directly by refined feature, while the predicates are predicted from the union features refined by the structures of BiLSTM [57], BiTreeLSTM [37], GRU [10], or GNN [27].

Nevertheless, most models are still trained on the whole dataset at one time, making the task challenging. To address the long-tail effect and mixed-granularity annotation, we believe it is a better solution to disentangle semantic-confusing predicate classes by dealing with general relationships (e.g., "on", "has") and informative ones (e.g., "sitting on", "using") separately in different stages. In this training scenario, the model in its stage can only focus on a small fraction of predicates with relatively similar granularity and then congregate the knowledge from previous stages step by step.

Since the SGG task is similar to the human learning pattern, absorbing knowledge from coarse to fine, it is natural to model how humans learn. Given the model trained in the previous stage, the current training model needs to

gain the ability to do well in previous classes while learning how to deal with new classes. A naive way is to sample some images in the last stage for review. Nevertheless, this strategy is unsuitable for our situation since it reintroduces mixed-granularity predicates. Thus, for better knowledge consolidation, we adopt **Concept Reconstruction (CR)** and **Model Reconstruction (MR)** which will be further explained in Sec. 3.3 and Sec. 3.4. CR will be adopted to decrease the distance between the prediction logits produced by the two models. This process is similar to how human students imitate teachers to solve problems. Human brain cortical areas have different functional networks [14]. It is the same for the parameters in an SGG model. MR respects the hypothesis that different parameters in a model serve for different relationship recognition. As is shown in Fig. 2, in the second to N^{th} stage, the model will fit the prediction and parameters for knowledge passing. At the same time, the gradient and parameters' change will be stored for the online update of the empirical Fisher Information Matrix (Sec. 3.4) and importance scores (Sec. 3.4).

Does a deep network have parameter redundancy? Some earlier work did answer this question. Hinton *et al.* [18] came up with knowledge distillation for the first time to compress and transfer knowledge from a complicated model to a more compact one that is easier to deploy. Moreover, it was verified in [3] that a cumbersome model can be compressed into a compact one. We assume that the whole model comprises many parameters based on these works. These parameters or "activated" neurons in an SGG network should target specific predicate classes, as shown in Fig. 1. To verify this assumption, we compare the mean of importance scores, which will be further illustrated in Sec. 3.4 and verified in supplementary material from the experimental perspective, of all the parameters in each layer and find out that the values vary between different stages. This mechanism is similar to how the human brain works. Each region in the brain has its' own function and works together to finish complex tasks [14]. After learning through all stages, the model will classify all classes with fine-grained preference.

The total loss for the HML framework is given by:

$$\ell = \ell_{new} + \ell_{CR} + \lambda \ell_{MR}, \quad (3)$$

where λ is a hyper-parameter. ℓ_{CR} is Concept Reconstruction loss, and ℓ_{MR} is Model Reconstruction loss.

ℓ_{new} in Eq. (3) is Class-Balance loss [8] and used to learn current stage predicates:

$$\ell_{new} = -W_B \sum_{i=1}^C y_i \log \frac{e^{z_i}}{\sum_{j=1}^C e^{z_j}}, \quad (4)$$

where z represents the model's output, y is the one-hot ground-truth label vector, and C is the number of predicate classes. $W_B = \frac{1-\gamma}{1-\gamma^{n_i}}$, γ denotes the hyper-parameter that represents the sample domain, and n_i denotes number of predicate i in current stage.

3.3 Concept Reconstruction Loss

In order to prevent activation drift [3], CR is applied. It is an implicit way to keep the prediction of previous predicates from drifting too much. Thus, we need to find the distance between two predictions from different stages of the same visual relationship and reduce it. The CR loss can be represented by:

$$\ell_{CR}(X_n, Z_n) = \frac{\sum_{i=1}^{N_n} \sum_{j=1}^{C_{old}} (\text{Softmax}(x_n^{ij}) - \text{Softmax}(z_n^{ij}))^2}{N_n}, \quad (5)$$

where x_n^i and z_n^i are the output logits vector for the prediction. N_n is the number of outputs. We choose L2 distance [2] as the distance metric. Compared with the traditional loss function such as L1 loss and cross-entropy loss with the soft label, L2 loss is a stronger constraint but is sensitive to outliers. Fortunately, since the training process is coarse to fine, the representations will not drastically deviate, making L2 loss practical. With the consideration mentioned earlier, L2 distance is used in CR, receiving better performance in experiments. This is also verified in [60]. More ablation results of CR can be found in the supplementary material.

In a word, CR is used to help the current model learn how to make the same prediction as the previous model.

3.4 Model Reconstruction Loss

The parameters of the model determine the ability to recognize visual relationships. Thus, it is a more straightforward way to learn directly from parameters. A feasible solution is to determine which parameters are crucial in the previous stage classification and fit them with greater attention in the following stage.

KL-divergence is a mathematical statistics measure of how a probability distribution is different from another one [25]. KL-divergence, denoted as in the form of $D_{KL}(p_\theta || p_{\theta+\Delta\theta})$, can also be used to calculate the difference of the conditional likelihood between a model at θ and $\theta + \Delta\theta$. Since changes of parameters are subtle (i.e., $\Delta\theta \rightarrow 0$) during the optimization, we will get the second-order of Taylor approximation of KL-divergence, which is also the distance in Riemannian manifold induced by Fisher Information Matrix [26] and can be written as $D_{KL}(p_\theta || p_{\theta+\Delta\theta}) \approx \frac{1}{2} \Delta\theta^\top F_\theta \Delta\theta$, where the F_θ is known as empirical Fisher Information Matrix [35] at θ and the approximate will be proved in supplementary material, is defined as:

$$F_\theta = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim D} \left[\left(\frac{\partial \log p_\theta(\mathbf{y}|\mathbf{x})}{\partial \theta} \right) \left(\frac{\partial \log p_\theta(\mathbf{y}|\mathbf{x})}{\partial \theta} \right)^\top \right], \quad (6)$$

where D is the dataset and $p_\theta(y|x)$ is the log-likelihood. However, in practice, if a model has P parameters, it means $F_\theta \in R^{P \times P}$, and it is computationally expensive. To solve this, we compromise and assume parameters are all independent, making F_θ diagonal. Then the approximation of KL-divergence looks like:

$$D_{KL}(p_\theta || p_{\theta+\Delta\theta}) \approx \frac{1}{2} \sum_{i=1}^P F_{\theta_i} \Delta\theta_i^2, \quad (7)$$

where θ_i is the i^{th} parameters of the model and P is the total number of it.

F_θ will be updated in each iteration, following the rule:

$$F_\theta^t = \frac{F_\theta^t + (t-1) \cdot F_\theta^{t-1}}{t}, \quad (8)$$

where t is the number of iterations.

Although the empirical Fisher Information Matrix captures static information of the model, it fails to capture the influence of each parameter during optimization in each stage. Thus, we adopt the method in [58] to search for essential parameters. Intuitively, if the value of a parameter changes a little in a single step, but it contributes a lot to the decrease of the loss, we think it is essential, at least for the current task. So, the importance of a parameter during an interval (from t to Δt) can be represented as:

$$\Omega_{raw}(\theta_i) = \sum_t^{t+\Delta t} \frac{\Delta \ell_t^{t+1}(\theta_i)}{\frac{1}{2} F_{\theta_i}^t (\theta_i(t+1) - \theta_i(t))^2 + \epsilon}, \quad (9)$$

$$\Omega_t^{t+\Delta t}(\theta_i) = \sigma \left(\log_{10} \frac{P \times \Omega_{raw}(\theta_i)}{\sum_{i=1}^P \Omega_{raw}(\theta_i)} \right), \quad (10)$$

where σ is the sigmoid function, numerator $\Delta \ell_t^{t+1}(\theta_i)$ is the change of loss caused by θ_i in one step, $\epsilon > 0$ aims to avoid the change of loss $\theta_i(t+1) - \theta_i(t) = 0$, and the denominator is the KL-divergence of θ_i between t and $t+1$.

To be more specific, $\Delta \ell_t^{t+1}(\theta_i)$ represents how much contribution does θ_i make to decrease the loss. Since the optimization trajectory is hard to track, to find the change in loss caused by θ_i , we need to figure out a way to split the overall loss form into the sum of each parameter's contributions. The solution is a first-order Taylor approximation:

$$\ell(\theta(t+1)) - \ell(\theta(t)) \approx - \sum_{i=1}^P \sum_{t=t}^{t+1} \frac{\partial \ell}{\partial \theta_i} (\theta_i(t+1) - \theta_i(t)) = - \sum_{i=1}^P \Delta \ell_t^{t+1}(\theta_i), \quad (11)$$

where $\frac{\partial \ell}{\partial \theta_i}$ is the gradient of θ_i and $\theta_i(t+1) - \theta_i(t)$ is the value change of θ_i during a single step. If $\ell(\theta(t+1)) - \ell(\theta(t)) > 0$, we set $\Delta \ell_t^{t+1}(\theta_i)$ to be 0, since we consider only when the loss become smaller, a step of optimization can be regarded as effective.

The empirical Fisher Information Matrix is used twice. The first is to calculate the difference in probability distributions of two models in different stages, and the second is to find the changes of a model in a nearby iteration within a single stage.

So, after figuring out how important each parameter is, the MR loss can be written as:

$$\ell_{MR} = \frac{\sum_{i=1}^P \left(F_{\theta_i}^{k-1} + (\Omega_{t_0}^{t_0+\Delta t})^{k-1}(\theta_i) \right) (\theta_i^k - \theta_i^{k-1})^2}{P}, \quad (12)$$

where P is the number of parameters for relationship prediction in the model and k represents the current stage. $F_{\theta_i}^{k-1}$ and $(\Omega_{t_0}^{t_0+\Delta t})^{k-1}$ are both calculated in previous stage.

Fisher Information Matrix F_{θ} and importance scores $\Omega_{t_0}^{t_0+\Delta t}$ are used to represent the importance of parameters from static and dynamic perspectives, respectively.

4 Experiment

4.1 Dataset and Model

Dataset. In the SGG task, we choose Visual Genome (VG) [24] as the dataset for both training and evaluation. It comprises 75k object categories and 40k predicate categories. However, due to the scarcity of over 90% predicates are less than ten instances, we applied the widely accepted split in [6,37,57], using the 150 highest frequency objects categories and 50 predicate categories. The training set is set to be 70%, and the testing set is 30%, with 5k images from the training set for finetuning. [42].

Model. We evaluate HML framework on three models and follow the setting in [41]: MOTIFS [57], Transformer [12,45], and VCTree [37].

4.2 Evaluation

Sub-Tasks: (1) **Predicate Classification:** given images, object bounding boxes, and object labels, predicting the relationship labels between objects. (2) **Scene Graph Classification:** given images and object bounding boxes, predicting object labels and relationship labels between objects. (3) **Scene Graph Detection:** localizing objects, recognizing objects, and predicting their relationships directly from images.

Relationship Recall. We choose **Mean Recall@K (mR@K)** [43] as a metric to evaluate the performance of SGG models. As is shown in [42], regular Recall@K (R@K) will lead to the reporting bias due to the imbalance that lies in the data (e.g., a model that only correctly classifies the top 5 frequent predicates can reach 75% of Recall@100). Thus, we introduce **Mean@K** which calculates the average score of all three sub-tasks R@K and mR@K under identical K. Mean@K is a metric to evaluate overall performance on both R@K and mR@K. We will further explain the reason and necessity of using this metric in the supplementary material.

4.3 Implementation Details

Object Detector. We pre-train a Faster-RCNN with ResNeXt-101-FPN [49] and freeze the previously trained weight during the SGG training period. The final detector reached 28 mAP on the VG test set.

Relationship Predictor. The backbone of baseline models is replaced with an identical one, and hierarchical trained in the same setting. We set the batch size

Model	Framework	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
		mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
Transformer [45]	Baseline	14.1	17.9	19.4	8.2	10.1	10.8	6.3	8.5	10.1
	CogTree [56]	22.9	28.4	31.0	13.0	15.7	16.7	7.9	11.1	12.7
	BPL+SA [16]	26.7	31.9	34.2	15.7	18.5	19.4	11.4	14.8	17.1
	HML(Ours)	27.4	33.3	35.9	15.7	19.1	20.4	11.4	15.0	17.7
MOTIFS [57]	Baseline	12.5	15.9	17.2	7.4	9.1	9.7	5.3	7.3	8.6
	EBM [39]	14.2	18.0	19.5	8.2	10.2	11.0	5.7	7.7	9.3
	SG [23]	14.5	18.5	20.2	8.9	11.2	12.1	6.4	8.3	9.2
	TDE [42]	18.5	25.5	29.1	9.8	13.1	14.9	5.8	8.2	9.8
	CogTree [56]	20.9	26.4	29.0	12.1	14.9	16.1	7.9	10.4	11.8
	DLFE [7]	22.1	26.9	28.8	12.8	15.2	15.9	8.6	11.7	13.8
	BPL+SA [16]	24.8	29.7	31.7	14.0	16.5	17.5	10.7	13.5	15.6
	GCL [11]	30.5	36.1	38.2	18.0	20.8	21.8	12.9	16.8	19.3
	HML(Ours)	30.1	36.3	38.7	17.1	20.8	22.1	10.8	14.6	17.3
VCTree [43]	Baseline	13.4	16.8	18.1	8.5	10.5	11.2	5.9	8.2	9.6
	EBM [39]	14.2	18.2	19.7	10.4	12.5	13.5	5.7	7.7	9.1
	SG [23]	15.0	19.2	21.1	9.3	11.6	12.3	6.3	8.1	9.0
	TDE [42]	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
	CogTree [56]	22.0	27.6	29.7	15.4	18.8	19.9	7.8	10.4	12.1
	DLFE [7]	20.8	25.3	27.1	15.8	18.9	20.0	8.6	11.8	13.8
	BPL+SA [16]	26.2	30.6	32.6	17.2	20.1	21.2	10.6	13.5	15.7
	GCL [11]	31.4	37.1	39.1	19.5	22.5	23.5	11.9	15.2	17.5
	HML(Ours)	31.0	36.9	39.2	20.5	25.0	26.8	10.1	13.7	16.3

(a) Comparison between HML and various optimization frameworks.

Model+Framework	Predicate Classification			Scene Graph Classification		Scene Graph Detection		Mean@50/100
	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100		
MOTIFS-TDE [42]	25.5/29.1	46.2/51.4	13.1/14.9	27.7/29.9	8.2/9.8	16.9/20.3	22.9/25.9	
MOTIFS-DLFE [7]	26.9/28.8	52.5/54.2	15.2/15.9	32.3/33.1	11.7/13.8	25.4/29.4	27.3/29.2	
Transformer-CogTree [56]	28.4/31.0	38.4/39.7	15.7/16.7	22.9/23.4	11.1/12.7	19.5/21.7	22.7/24.2	
PCPL [52]	35.2/37.8	50.8/52.6	18.6/19.6	27.6/28.4	9.5/11.7	14.6/18.6	26.1/28.1	
DT2-ACBS [9]	35.9/39.7	23.3/25.6	24.8/ 27.5	16.2/17.6	22.0/24.4	15.0/16.3	22.9/25.2	
SHA-GCL [11]	41.6/44.1	35.1/37.2	23.0/24.3	22.8/23.9	17.9/20.9	14.9/18.2	25.9/28.1	
Transformer-HML(Ours)	33.3/35.9	45.6/47.8	19.1/20.4	22.5/23.8	15.0/17.7	15.4/18.6	25.2/27.4	
MOTIFS-HML(Ours)	36.3/38.7	47.1/49.1	20.8/22.1	26.1/27.4	14.6/17.3	17.6/21.1	27.1/29.3	
VCTree-HML(Ours)	36.9/39.2	47.0/48.8	25.0/26.8	27.0/28.4	13.7/16.3	17.6/21.0	27.9/30.1	

(b) A more comprehensive comparison between HML and various SOTAs.

Table 1: Result of Relationship Retrieval mR@K [43] and R@K.

to 12 and used SGD optimizer with an initial learning rate of 0.001, which will be decayed by 10 after the validation performance plateaus. The experiment was carried out on NVIDIA TITAN RTX GPUs.

Hierarchical Predicate Tree Construction. For Visual Genome (VG), the whole predicates are split into two disjoint subsets (tasks) following the construction in Sec. 3.1. The threshold is set to be $T_{SS} \in [0.65, 0.75]$ in Eq. (2). However, due to the small semantic variance within VG, this Predicate Tree Construction degenerates to simply separating frequent “on” or “has” with minority ones, similar to the separation in [11]. We believe a larger semantic variance dataset will need this kind of semantic information for more reasonable separation.

Hierarchical Training. In our experiment, two identical models will be trained separately in two stages. Moreover, the first model will be initialized randomly,

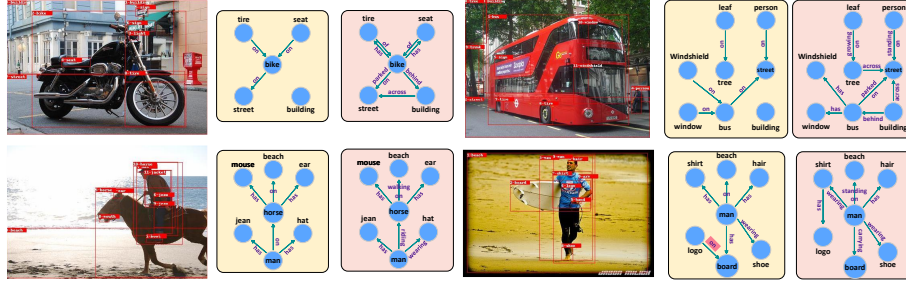


Fig. 3: **Qualitative Results.** The basic model generates yellow scene graphs, and the same basic model predicts pink ones under HML training.

and the previous stage models will not be used to initialize the following model. If not, the model will meet the intransigence [5] problem. Besides, we set the max iterations to 8000 and 16000 in the first and second stages. Nevertheless, models will usually converge around 8000 in both stages. The λ in an overall loss is set to 0.5 for the best performance, as is shown in Tab. 3a.

4.4 Quantitative Study

In Tab. 1a, our HML framework can consistently outperform the existing frameworks under various fundamental models. By applying the HML framework to different models, we notice that models will drastically improve the Mean Recall for all three sub-tasks. The improvement scale is similar to different models such as Motif, Transformer, and VCTree. In Tab. 1b, we compare HML with various SOTAs, which report R@K and mR@K simultaneously. After calculating Mean@K, it turns out that HML can make the fine-grained prediction (demonstrated by mR@K) and keep clear boundaries of general predicates (demonstrated by R@K).

In Tab. 2, we notice a vital decrease in R@K. After analyzing each predicate, we figure out that the decline mainly comes from the drop of general predicates such as “on” and “has”. After HML training, the mR@100 of “on” and “has” dropped from 79.98 to 31.01 and 81.22 to 58.34. The model’s preference for fine-grained prediction causes this decrease. Thus, we replace the fine-grained one with a general one and recalculate R@100, and the results of “on” and “has” re-bounce to be 71.72 and 86.91, respectively. Works like [7,9,42,52,56] compared in Tab. 1b can also make relatively fine-grained prediction (i.e., relatively high mR@K) but all suffer from decrease of R@K partially due to the reason mentioned above.

4.5 Qualitative Study

We visualize the qualitative result generated by original MOTIFS and MOTIFS trained with the HML framework in Fig. 3. Compared with the original model,

Model	CR	MR	Predicate Classification					
			mR@20	mR@50	mR@100	R@20	R@50	R@100
Transformer [45]			14.13	17.87	19.38	58.79	65.29	67.09
	✓		24.14	29.27	31.22	29.1	36.16	38.57
		✓	23.32	29.34	32.20	38.80	46.48	48.87
	✓	✓	27.35	33.25	35.85	38.81	45.61	47.78
MOTIFS [57]			12.54	15.89	17.19	59.12	65.45	67.20
	✓		24.69	30.00	32.79	33.92	41.34	43.95
		✓	21.56	27.43	30.05	44.30	51.87	54.14
	✓	✓	30.10	36.28	38.67	40.52	47.11	49.08
VCTree [43]			13.36	16.81	18.08	59.76	65.48	67.49
	✓		22.32	29.34	32.20	27.86	35.21	37.73
		✓	22.84	28.94	31.48	43.81	51.40	53.74
	✓	✓	31.04	36.90	39.21	40.28	46.47	48.36

Table 2: **Ablation on CR and MR.** We explore the functionality of CR loss and MR loss.

the model trained under HML will make informative predictions such as “parked on”, “growing on”, “standing on”, “riding”, and “walking on” instead of the general one “on”. Also, our model will tend to make fine-grained predictions such as “wearing” and “carrying” instead of “has”. Besides, since we train a model hierarchically, the model will obtain the ability to capture tail part predicates such as “building-across-street” and position predicates such as “building-behind-bus” and “tree-behind-bus”. Qualitative results with three stages are shown in the supplementary material.

4.6 Ablation Studies

CR and MR. We further explore the contributions of each term to our overall loss. In the SGG task, Recall@K and Mean Recall@K [43] restrict mutually with each other. Recall@K represents how well the model performs for the predicate classes’ head part. Thus, it reflects how well a model can imitate the previous model. On the contrary, Mean Recall@K [43] evaluates the model’s overall performance. Suppose we want to figure out the functionality of the knowledge consolidation term in the loss. In that case, it is reasonable to adopt Recall@K since two terms of knowledge reconstruction aim to prevent the model from forgetting previous knowledge. According to Tab. 2, if we add CR and MR separately, mR@K will get constant improvement. However, only when CR and MR are used simultaneously will we get the highest mR@K and prevent R@K from dropping too much. Also, after comparing the second and third row of each model on R@K, it is obvious that MR is a more powerful constraint than CR.

Layer Number. The number of layers (i.e., stage) depends on how many top-K frequent predicates we pick up after clustering. We conduct experiments in

layer	Predicates Classification			time (hr)	λ	Predicates Classification		
	mR@20	mR@50	mR@100			mR@20	mR@50	mR@100
1	12.54	15.89	17.19	17.85	0.00	25.24	31.95	34.44
2	30.10	36.28	38.67	29.43	0.25	30.97	35.91	37.88
3	25.24	31.95	34.44	44.65	0.50	30.10	36.28	38.67
4	15.66	21.96	25.32	60.13	0.75	29.73	34.36	36.79
					1.00	29.20	33.47	35.69

(a) Ablation of Layer Number

(b) Ablation of λ Table 3: **Ablation on MOTIFS.** We explore different numbers of layers and λ with MOTIFS on the performance of Mean Recall@K.

Tab. 3a on different layers. We figured out that 2 is suitable for the VG dataset, mainly due to the small number of predicate classes and limited granularity variance. HML training indeed needs more time to complete training during multi-stage training. Nevertheless, the increase ratio (125%) of model performance is way more significant than the one (65%) of training time. More time analysis will be shown in the supplementary material. All experiments were carried out on one identical GPU.

Hyperparameter λ . In order to figure out the effect of λ on the performance of the model, we set 5 values in ablation to $\lambda \in \{0, 0.25, 0.50, 0.75, 1.00\}$ in Tab. 3b. λ represents how much information will be passed down to the next stage. If λ is too high, the new model will stick to the original classes without learning new ones. In contrast, low λ can not guarantee effective information passing. Based on our experiment, $\lambda = 0.5$ is suitable for the HML framework on VG.

5 Conclusion

We propose a general framework to enable SGG models to make fine-grained predictions. In addition to the objective long-tail effect in the dataset, we uncover mixed-granularity predicates caused by subjective human annotation. The similarity between the human hierarchical learning pattern and the SGG problem is obvious under this condition. Based on that, we designed the HML framework with two new constraints (i.e., CR and MR) for efficient training. We observe that the HML framework can improve performance compared to the traditional training fashion models and achieves new state-of-the-art.

Acknowledgments This work was partly supported by the National Natural Science Foundation of China under Grant 41971284; the Fundamental Research Funds for the Central Universities under Grant 2042022kf1201; Wuhan University-Huawei Geoinformatics Innovation Laboratory. We sincerely thank our reviewers and ACs for providing insightful suggestions.

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: CVPR. pp. 4971–4980 (2018) [1](#)
2. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: NIPS. pp. 2654–2662 (2014) [8](#)
3. Buciluă, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: KDD. pp. 535–541 (2006) [7](#), [8](#)
4. Cao, D., Zhu, X., Huang, X., Guo, J., Lei, Z.: Domain balancing: Face recognition on long-tailed domains. In: CVPR. pp. 5671–5679 (2020) [4](#)
5. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: ECCV. pp. 532–547 (2018) [12](#)
6. Chen, L., Zhang, H., Xiao, J., He, X., Pu, S., Chang, S.F.: Counterfactual critic multi-agent training for scene graph generation. In: ICCV. pp. 4613–4623 (2019) [10](#)
7. Chiou, M.J., Ding, H., Yan, H., Wang, C., Zimmermann, R., Feng, J.: Recovering the unbiased scene graphs from the biased ones. In: ACMMM. pp. 1581–1590 (2021) [2](#), [4](#), [11](#), [12](#)
8. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR. pp. 9268–9277 (2019) [4](#), [7](#)
9. Desai, A., Wu, T.Y., Tripathi, S., Vasconcelos, N.: Learning of visual relations: The devil is in the tails. In: ICCV. pp. 15404–15413 (2021) [4](#), [11](#), [12](#)
10. Dhingra, N., Ritter, F., Kunz, A.: Bgt-net: Bidirectional gru transformer network for scene graph generation. In: CVPR. pp. 2150–2159 (2021) [6](#)
11. Dong, X., Gan, T., Song, X., Wu, J., Cheng, Y., Nie, L.: Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In: CVPR. pp. 19427–19436 (2022) [4](#), [11](#)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) [10](#)
13. Feng, C., Zhong, Y., Huang, W.: Exploring classification equilibrium in long-tailed object detection. In: ICCV. pp. 3417–3426 (2021) [4](#)
14. Genon, S., Reid, A., Langner, R., Amunts, K., Eickhoff, S.B.: How to characterize the function of a brain region. *Trends in cognitive sciences* **22**(4), 350–364 (2018) [3](#), [7](#)
15. Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. In: ICLR (2014) [4](#)
16. Guo, Y., Gao, L., Wang, X., Hu, Y., Xu, X., Lu, X., Shen, H.T., Song, J.: From general to specific: Informative scene graph generation via balance adjustment. In: ICCV. pp. 16383–16392 (2021) [2](#), [4](#), [11](#)
17. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: ECCV. pp. 771–787 (2018) [1](#)
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. *CoRR* (2015) [7](#)

19. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR. pp. 6700–6709 (2019) [3](#)
20. Hung, Z., Mallya, A., Lazebnik, S.: Contextual translation embedding for visual relationship detection and scene graph generation. TPAMI **43**(11), 3820–3832 (Nov 2021) [4](#)
21. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. ICLR (2020) [4](#)
22. Khan, S., Hayat, M., Zamir, S.W., Shen, J., Shao, L.: Striking the right balance with uncertainty. In: CVPR. pp. 103–112 (2019) [4](#)
23. Khandelwal, S., Suhail, M., Sigal, L.: Segmentation-grounded scene graph generation. In: ICCV. pp. 15879–15889 (2021) [11](#)
24. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV **123**(1), 32–73 (2017) [10](#)
25. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951) [8](#)
26. Lee, J.M.: Riemannian manifolds: an introduction to curvature, vol. 176. Springer Science & Business Media (2006) [8](#)
27. Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: CVPR. pp. 11109–11119 (2021) [4](#), [6](#)
28. Li, T., Wang, L., Wu, G.: Self supervision to distillation for long-tailed visual recognition. In: ICCV. pp. 630–639 (2021) [4](#)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014) [2](#), [4](#)
30. Liu, J., Sun, Y., Han, C., Dou, Z., Li, W.: Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In: CVPR. pp. 2970–2979 (2020) [4](#)
31. Manjunatha, V., Saini, N., Davis, L.S.: Explicit bias discovery in visual question answering models. In: CVPR. pp. 9562–9571 (2019) [1](#)
32. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, pp. 109–165 (1989) [4](#)
33. Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., Joulin, A.: Advances in pre-training distributed word representations. In: LREC (2018) [5](#)
34. Misra, I., Lawrence Zitnick, C., Mitchell, M., Girshick, R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In: CVPR. pp. 2930–2939 (2016) [5](#)
35. Pascanu, R., Bengio, Y.: Revisiting natural gradient for deep networks. ICLR (2014) [8](#)
36. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: ICCV. pp. 1981–1990 (2019) [4](#)
37. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: CVPR (July 2017) [6](#), [10](#)
38. Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. In: NIPS. pp. 4175–4186 (2020) [4](#)
39. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.: Energy-based learning for scene graph generation. In: CVPR. pp. 13936–13945 (2021) [2](#), [4](#), [11](#)

40. Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: CVPR. pp. 11662–11671 (2020) [4](#)
41. Tang, K.: A scene graph generation codebase in pytorch (2020), <https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch> [10](#)
42. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: CVPR. pp. 3716–3725 (2020) [2, 4, 10, 11, 12](#)
43. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: CVPR. pp. 6619–6628 (2019) [1, 4, 10, 11, 13](#)
44. Tao, L., Mi, L., Li, N., Cheng, X., Hu, Y., Chen, Z.: Predicate correlation learning for scene graph generation. TIP (2022) [3](#)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS. pp. 5998–6008 (2017) [10, 11, 13](#)
46. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph matching for relationship-aware image-text retrieval. In: WACV. pp. 1508–1517 (2020) [3](#)
47. Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J.: The devil is in classification: A simple framework for long-tail instance segmentation. In: ECCV. pp. 728–744 (2021) [4](#)
48. Wang, Y., Ramanan, D., Hebert, M.: Learning to model the tail. In: NIPS. pp. 7029–7039 (2017) [4](#)
49. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. pp. 1492–1500 (2017) [10](#)
50. Xiong, S., Huang, W., Duan, P.: Knowledge graph embedding via relation paths and dynamic mapping matrix. In: Advances in Conceptual Modeling. pp. 106–118 (2018) [4](#)
51. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR. pp. 5410–5419 (2017) [1, 3, 4](#)
52. Yan, S., Shen, C., Jin, Z., Huang, J., Jiang, R., Chen, Y., Hua, X.S.: Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In: ACM-MM. pp. 265–273 (2020) [11, 12](#)
53. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: ECCV. pp. 670–685 (2018) [1](#)
54. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: CVPR. pp. 10685–10694 (2019) [1, 3](#)
55. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Feature transfer learning for face recognition with under-represented data. In: CVPR. pp. 5704–5713 (2019) [4](#)
56. Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q.: Cogtree: Cognition tree loss for unbiased scene graph generation. In: IJCAI. pp. 1274–1280 (2020) [4, 11, 12](#)
57. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR. pp. 5831–5840 (2018) [1, 3, 4, 5, 6, 10, 11, 13](#)
58. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: ICML. pp. 3987–3995 (2017) [9](#)
59. Zhang, H., Kyaw, Z., Chang, S., Chua, T.: Visual translation embedding network for visual relation detection. In: CVPR. pp. 3107–3115 (2017) [4](#)
60. Zhang, J., Zhang, J., Ghosh, S., Li, D., Tasci, S., Heck, L., Zhang, H., Kuo, C.C.J.: Class-incremental learning via deep model consolidation. In: WACV. pp. 1131–1140 (2020) [8](#)
61. Zhang, X., Wu, Z., Weng, Z., Fu, H., Chen, J., Jiang, Y.G., Davis, L.S.: Videolt: Large-scale long-tailed video recognition. In: ICCV. pp. 7960–7969 (2021) [4](#)