This ACCV 2022 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2022
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Continuous Self-Study: Scene Graph Generation with Self-Knowledge Distillation and Spatial Augmentation*

Yuan Lv, Yajing Xu, Shusen Wang, Yingjian Ma, and Dengke Wang

Beijing University of Posts and Telecommunications
{lvyuan,xyj, ShusenW, wangdk}@bupt.edu.cn

Abstract. As an extension of visual detection tasks, scene graph generation (SGG) has drawn increasing attention with the achievement of complex image understanding. However, it still faces two challenges: one is the distinguishing of objects with high visual similarity, the other is the discriminating of relationships with long-tailed bias. In this paper, we propose a Continuous Self-Study model (CSS) with self-knowledge distillation and spatial augmentation to refine the detection of hard samples. We design a long-term memory structure for CSS to learn its own behavior with the context feature, which can perceive the hard sample of itself and focus more on similar targets in different scenes. Meanwhile, a fine-grained relative position encoding method is adopted to augment spatial features and supplement relationship information. On the Visual Genome benchmark, experiments show that the proposed CSS achieves obvious improvements over the previous state-of-the-art methods. Our code is available at https://github.com/LINYE1998/Continuous_Self_Study.

1 Introduction

Scene graph [1,2] structure is a medium bridging the image and the text [3,4]. It is comprised by the detection of a list of ⟨*subject-predicate-object*⟩ triplets [5] to describe the objects and their relationships in an image. With feature augmentation by the extraction of context information [1,6,7] and the introduction of external semantic knowledge [8,9,10], it can not only improve the accuracy of classification in upstream tasks, such as object detection [11] and visual relationship detection [12,13,10], but also provide a more comprehensive and specific structure for its downstream visual understanding tasks [14,15], including image retrieval [16], visual question answering [17,18] and image captioning [19], thus has been drawing increasing attention.

To generate high-quality scene graphs, multifarious scene graph generation (SGG) methods [1,2,6,20,21] have been proposed to optimize the prediction of objects and relations. It can be mainly classified as the traditional SGG types [1,6] and the unbiased SGG types [20,22]. Both approaches refine the targets by passing visual or semantic messages with the extraction of context

* Supported by National Natural Science Foundation of China.

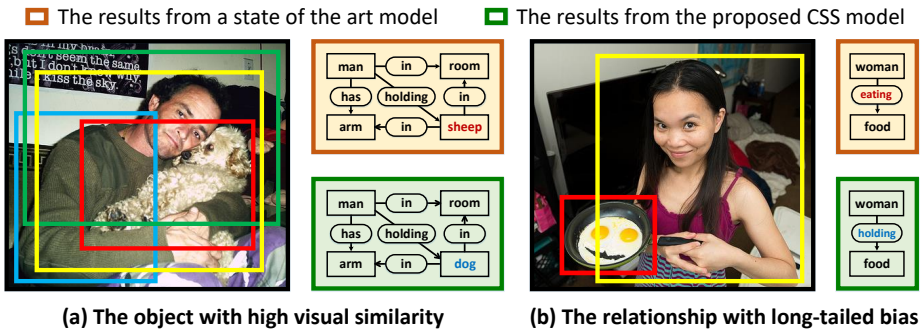


Fig. 1. Examples of the two challenges for SGG task. (1) The distinguish of the unclear target, e.g., the object *dog* is predicted as *sheep* wrongly in (a). (2) The discriminating of the long-tailed relationship, e.g., the *holding* is misclassified as *eating* in (b), due to the common sense bias [2] from ⟨*woman-eating-food*⟩.

information [1,7]. Differently, the first types focus on a better feature extraction network [2,23,24] to mine useful information from more perspectives, while the second types concentrate on the debiasing work [20,21,25] to recall more semantic relationships [2] and obtain a more balancing result for the application of downstream tasks [20]. Although the previous methods have promising improvement in performance, most of them suffer from the limitations of existing SGG datasets [26]: the inadequate training data with hard sample, and the unbalanced distribution of the long-tailed relation.

The atypical objects with high visual similarity are always hard to be distinguished. For example, in the red box in Fig. 1(a), the *dog* is identified as *sheep* by a state-of-the-art model. While the *dog* is ambiguous and difficult to be distinguished, it's easy to recognize it by inference with the context information. To extract the scene information from the image, numerous researchers [2,6,7,10,27] struggle for better feature extraction networks [20]. However, it's still difficult to understand the scene and focus on the hard samples under dozens of predicted objects and biased relation of square growth. For human beings, how do we think when observing objects that are difficult to distinguish? Focusing on the unclear targets, we usually realize that we are confused and list several alternative possibilities, and then make the judgment in combination with the scene information. Inspired by the recent knowledge distillation work [28,29,30,31] , we want to enable the network to perceive the hard samples of itself and distinguish them with the supplement of scene context information.

Meanwhile, the recent research focuses on the debiasing work to balance the results from long-tailed bias. However, with the increase of the mean recall (mR@K) [23] among each predicate, most of the debiasing methods cause an unacceptable decline in Recall (R@K). As illustrated in Fig. 2, the unbiased SGG approaches choose a preference for the relationships with similar semantics, but finitely to predict them more accurately. It means that the upper limit of the

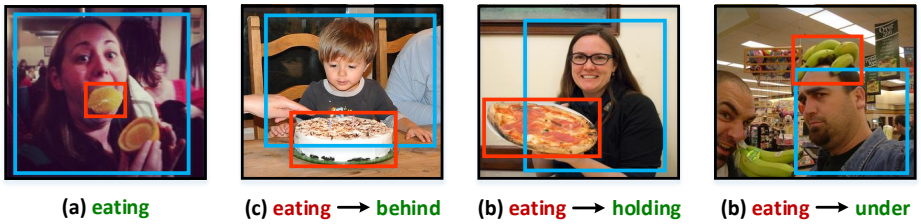


Fig. 2. Four typical cases of the long-tailed bias. With the subject *human* and the object *food*, different relations are more likely to be predicted as eating. By supplementing the relative position information, the relationship can be predicted correctly, e.g., from eating to *behind* in (b), *holding* in (c), and *under* in (d).

predicates' prediction is limited by the information the neural networks extract. Hence, the unbiased SGG methods still need to optimize the feature extraction networks for more useful information. A large proportion of predicates have a high correlation with the spatial relations between their objects. Therefore, Both SGG types [2,12,20,22,25] and much more visual detection tasks [32,33] adopt position encoding to extract spatial information. The common position encoding cuts the image averagely into a set resolution and encodes each object separately. Nevertheless, it cannot extract the relative position information explicitly, while the relative position can provide more details between the object pairs. Besides, for small objects, it's hard to extract accurate spatial information. This motivates us to augment the spatial feature with fine-grained encoding.

Hence, for the distinguishing of objects with high visual similarity, we proposed a novel self-distillation method: Continuous Self-Study (CSS) for SGG model to learn from its own behavior with a real-time updated long-term memory structure. Focusing on the hard sample, CSS transfers the detection task from the prediction of objects to the distinguishing of similar targets. Moreover, for the discriminating of relationships with long-tailed bias, we propose a spatial augmentation (SA) of the relative position to improve the spatial information from ambiguous and directionless to accurate and directional.

Our contributions can be summarized as follows:

- A Continuous Self-Study method for SGG models is proposed to learn self-behavior, so as to obtain better visual understanding and distinguish similar targets in complex scenes.
- A spatial augmentation method is designed for visual relation detection to effectively improve the recall (R@K) and mean recall (mR@K) among each predicate in unbiased SGG field.
- Experiments on the benchmark dataset show that our approach can improve on the state-of-the-art baseline.

2 Related Work

Scene Graph Generation. Scene graph [1,2] is a mid-connection [3,4,16] of visual domain and semantic domain, which drawn increasing attention with its refinement of visual detection tasks [5,11,12,13,34] and its potential value in several downstream visual reasoning [17,18,19,35,36,37] and visual understanding tasks [14,15]. The development of scene graph generation task can be divided into two stages. In the first stage, various methods [1,6,8,13,23,27,38,39,40] are proposed to explore multiple ways of the extraction of the feature. With the supplement of context feature [1,2,6,7] and the introduction of external language information [3,8,9,10], these methods access promising improvement of function and performance. However, This scene graph generation is far from practical, due to the biased SGG problem [3,20] with the long-tailed dataset.

In the second stage, multiple approaches are proposed to generate unbiased scene graph. Zellers *et al.* [2] firstly pointed out the bias problem of SGG and the followers [9,23,26] proposed the unbiased metric to evaluate SG with increased attention on tail relationships. Tang *et al.* [20] draw the counterfactual causality from the trained graph to infer the effect from the bias. Yu *et al.* [22] proposed a cognition tree loss to make the tail classes receive more attention in a coarse-to-fine mode. Guo *et al.* [21] tackled the bias problem with semantic adjustment and balanced predicate learning. Chiou *et al.* [25] used a dynamic label frequency estimation to balance the head and the tail data. However, the recent approaches struggle for the identification of tail predicates and focus on the promotion of the mean recall. With the improvement of mR@K, the recall of the head data got a severe drop, which made the SGG still far from practical.

Knowledge Distillation. Knowledge distillation [41,42,43,44,45] is a method of extraction, generalization, and transmission of knowledge. By transferring the knowledge [42,43,44,46,47] of a complex pre-trained teacher network [42,43,46], a simple student network [29,31,41,48,49] can be trained effectively with the pseudo labels. To address the problem of confirmation bias [50] in pseudo-labeling, Pham *et al.* [51] trained the teacher along with the student and corrected the bias with the feedback of the student's performance. However, these traditional methods depended on a well-trained teacher network [28]. Several self-knowledge distillation methods [52,53,54,30,55] are proposed to reduce the necessity of training a large network. Nevertheless, because of the square growth relation [7] with the targets, it's still hard for SGG model to overcome the limited computing [28,41,56,57] which inversely optimizes the detection of hard samples. To this end, we distill the knowledge from the network with a memory structure that enables the network to study from its own behavior.

3 Methodology

As illustrated in Fig. 3, the CSS model consists of two parts: (1) the Self-Study module (SS) for object refinement. It retains its behavior information in a real-time updated memory *Memorandum*. The hard samples are perceived by distilling the knowledge from *Memorandum* and combining it with the detected

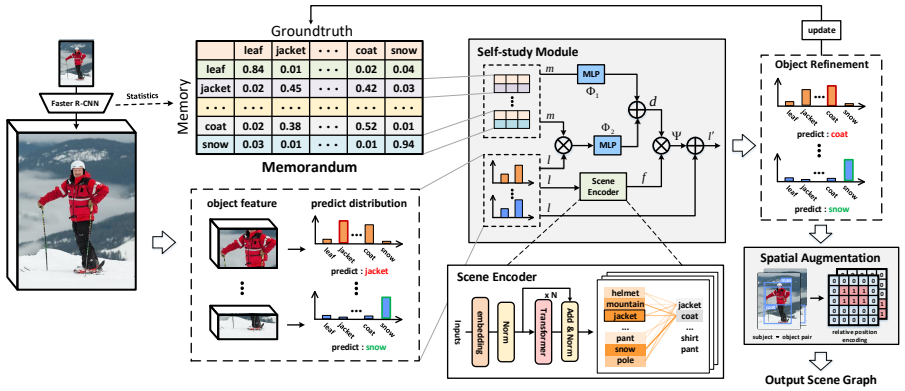


Fig. 3. Overview of the proposed Continuous Self-Study model. For the input image, the proposals are generated by a detector. With the object refinement by the self-study module, the hard sample **jacket** which is predicted wrongly by the detector can be refined to **coat** correctly. Then the relationship prediction is optimized with the spatial augmentation module. The output scene graph is generated with the combination of the predicted pair-wise objects and their detected relationships.

results. Then the hard samples are focused on to distinguished and refined with the supplement of the scene context. (2) the spatial augmentation module for relationship optimization. It embeds the spatial feature of the relative position through a fine-grained encoding with an explicit spatial constraint, which distinguishes the bidirectional relationships.

The input of the CSS are proposals generated by a detector. In order to describe better, the following definitions are given. For an input image I, We use a pre-trained Faster RCNN [58] as an underlying detector [2] to predict a set of region proposals $B = \{b_i\}$ and their corresponding detected object $O = \{o_i\}$. The proposal $b_i \in \mathbb{R}^4$ is represented by a bounding box $b_i = [x_i, y_i, w_i, h_i]$, where (x_i, y_i) are the coordinates of the box's top left corner, w_i and h_i are the width and height of the bounding box respectively. Meanwhile, the detector extracts a set of visual feature vector $V = \{v_i\}$ for each proposal b_i. With the feature vector, the Roi Box Head outputs a set of predicted vector $L = \{L_i\}$ which represents the per-class confidence distribution. In addition, the $C = \{c_i\}, i \in \{1, ..., R_c\}$ is the category set of the object, where R_c is the dimension.

3.1 Self-Study Module

Memorandum. We design a long-term memory structure, Memorandum, for CSS to retain its behavior information. The Memorandum $M \in \mathbb{R}^{R_c \times R_c}$ is a square matrix represented by a set of memory vectors $M = \{m_{i\cdot}\}$, where $m_{i\cdot}$ is the memory of the object class $c_i \in R_c$ which records the CSS's historical

predicted behavior. More specifically, the scalar m_{ij} represents the conditional probability $P(gt = c_j|pred = c_i)$ that the detector predicts the object as c_i while its ground truth is c_j, as illustrated in Fig. 3.

Intuitively, for a well-trained class c_w of CSS model, the memory vector $m_{w\cdot}$ will just activate at node m_{ww} with the rest of inactive nodes of $m_{w\cdot}$. On the contrary, for the indistinguishable classes, the memory vector will activate at pairwise even more nodes. Hence, the Memorandum structure can be regarded as a summary note organized by the CSS itself.

The CSS is trained with a two-stage strategy to avoid the difficulty of convergence caused by error accumulation at the beginning of training. In the first stage, the SGG network is trained without Memorandum until the network achieves the performance of the baseline. In the second stage, we initialize a heatmap $H \in \mathbb{R}^{R_C \times R_C}$ with the statistical matrix $S_0 = \{s_{ij}\} \in \mathbb{R}^{R_C \times R_C}$, where s_{ij} is the statistical quantity of the c_i predicted by CSS with the ground truth of c_j. Then, the heatmap is updated with the new statistics S_i each iteration, where S_i is the i-th statistical matrix of its iteration similar to S_0. Considering that the relevance between the current and historical state of the heatmap on CSS decreases over time, the historical data of each iteration is attenuated during the training process with the variant formula of Newton's law of cooling [59]:

$$T(t) = T(0)e^{-\alpha t} \tag{1}$$

where $T(0)$ and $T(t)$ are the temperature of time 0 and t respectively, and α is the attenuation factor. Eq. 1 can be regarded as a cooling process for an impulse response, which is widely used to calculate the heat of events in today's social network [60]. By summing the impulse responses after each iteration of training, the heatmap H at moment t can be calculated with Eq. 2:

$$H(t) = \sum_{i=0}^{t} S_i e^{-\alpha(t-i)} \tag{2}$$

With the increasing decay over time, the $H(t)$ can reflect the behavior of CSS with an appropriate cycle, which addresses the accumulation of too much historical behavior. To simplify the calculating process of $H(t)$, Eq. 2 can be further converted by making a difference between $H(t+1)$ and $H(t)$:

$$H(t+1) - H(t) = \sum_{i=0}^{t+1} S_i e^{-\alpha(t+1-i)} - \sum_{i=0}^{t} S_i e^{-\alpha(t-i)} \tag{3}$$

which can be simplified as:

$$H(t+1) = e^{-\alpha} H(t) + S_{t+1} \tag{4}$$

Eq. 4 will be derived in detail in the supplementary materials. The heatmap can be updated iteratively only through the statement of the last moment and the statistical matrix of current moment by Eq. 4. Then the Memorandum can be calculated with H:

$$M_i = LineNorm(H(i)) \tag{5}$$

where $LineNorm$ is a function that normalizes each row of H separately. By repeatedly distilling and updating knowledge from Memorandum, CSS can continuously study from its own behavior and finally get a well-trained network with an ideal Memorandum, which is a dynamic balanced diagonal matrix.

Knowledge Distillation. This subsection is to perceive and refine the hard samples with the behavior information distilling from the Memorandum and the supplement of the scene information. As illustrated in Fig. 3, each object O_i can be predicted with a pre-labeled c_i by getting the maximum value of the confidence vector l_i. The memory m_{c_i}.is drawn out from the Memorandum with the pre-predicted c_i.Then we use a perceiving layer to get the focus feature d_i which focus on the hard sample:

$$d_i = \Phi_1(l_i \odot m_{c_i}) + \alpha\Phi_2(m_{c_i}) \tag{6}$$

where Φ_1 and Φ_2 are multi-layer perceptron, α is a balance hyperparameter, and $\odot$ denotes the element-wise product. The focus feature can be regarded as a confusion vector with the confusion degree of each class. For the confidence vector l_i and the drawn-out memory m_{c_i} with only one activate node, the confusion degree will be very low for all nodes of d_i so that the CSS can decrease the correction with the scene feature for o_i. On the contrary, if l_i or m_{c_i} has two or more activate nodes, the confusion degree will be high between the class relative to the activate nodes, which will increase the refinement with the scene attention feature $F = \{f_i\}$.

As shown in the bottom of Fig. 3, the scene attention feature is extracted with the scene encoder network. In this work, we embed and normalize the l_i first, and then we use N transformer-based Encoder which connected end to end to adaptively gather contextual information for a certain object. The f_i can be regarded as an inference that predicts the probability distribution of c_i type object under a certain scenario.

Object Refinement. The objects are refined with the combination and fusion of the focus feature and the scene attention feature. To avoid the deviation from the image, the refinement needs to be constraint with the original visual information. Through the supplement of the original confidence distribution l, the predicted label of the object is refined by:

$$l'_i = \text{softmax}(l_i + \beta\Psi(f_i \odot d_i)) \tag{7}$$

where l'_i is the confidence distribution of the object after the refinement, β is a balance hyperparameter, and Ψ is a projection function. We denote $\beta\Psi(f_i \odot d_i)$ as $\tilde{l}$ with the statistical of the refined predicts, This Self-Study structure refine the distribution of classification probability by:

$$P(L') = P(L|V) + P(\tilde{L}|D, O_1, ..., O_n) \cdot P(D|L, M) \tag{8}$$

where $\tilde{L} = \tilde{l}_i$ is the refinement of the object, $L' = \{l'_i\}$ is the final output prediction distribution, and $D = d_i$ is the focus feature set. Eq. 8 embodies the essence of the self-study method. $P(D|L, M)$ is the knowledge distilled from

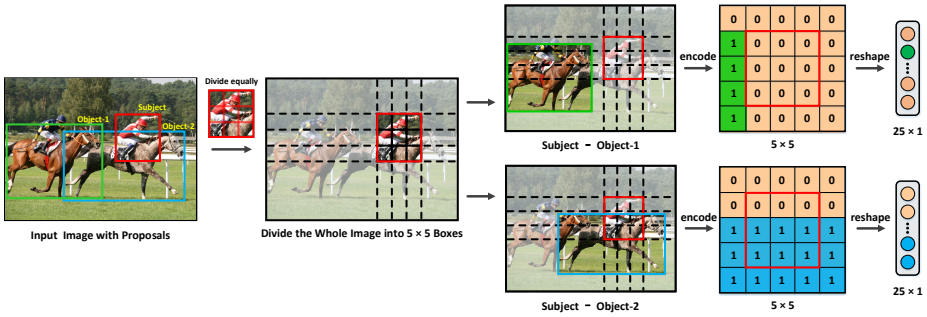


Fig. 4. The fine-grained encoding of the relative position. For the pairwise targets, the subject box is divided equally into nine parts. The whole image can be divided into twenty-five little boxes with the split lines obtained in the previous step. Then these boxes are encoded into a matrix, and the boxes which intersect with the object box are encoded as one. Finally, The matrixes are reshaped as a vector.

the CSS. For hard samples with high $P(D|L, M)$, CSS will focus more on scene context between the confusion classes with less attention from visual features. It is a positive feedback process to continuous self-study because the network will be refined by learning of the Memorandum, while the Memorandum will transform to a better distribution with the better performance of CSS.

3.2 Spatial Augmentation Module

To enhance the spatial constraint for the prediction of relations, we augment the relation feature with a fine-grained relative position spatial encoding. For better description, we use b_s and b_o to distinguish the bounding box of subject and object. As illustrated Fig. 4, for each triplet $\langle$*subject -predicate-object*$\rangle$, $b_s = [x, y, w, h]$ is divided equally into nine little boxes. Then, b_s is expanded into a larger box b_u, which obtained the whole view on the image I. With the nine boxes inside b_s and sixteen boxes outside b_s, the image I can be divided into twenty-five region $b_u = \{z_{ij}\}$, where z_{ij} can be represented by:

$$z_{ij} = [x_{ij}, y_{ij}, w_{ij}, h_{ij}], \quad i, j = 0, ..., 4 \tag{9}$$

where x_{ij}, w_{ij} can be further represented as:

$$x_{ij} = \begin{cases} 0, & j = 0 \\ x - (1 - j)\frac{1}{3}w, & j = 1, ..., 4 \end{cases} \tag{10}$$

$$w_{ij} = \begin{cases} x, & i = 0 \\ \frac{1}{3}w, & i = 1, 2, 3 \\ w_I - x - w, & i = 4 \end{cases} \tag{11}$$

where w_I is the width of the image. Meanwhile, y_{ij}, h_{ij} can be calculated in the same way with Eq. 10 and Eq. 11 respectively. The set Q_b is defined as the region within the bounding box b, and b_o is defined as the bounding box of the triplet's object. The spatial embedding of $\langle b_s, bo \rangle$ is encoded by a boolean matrix $H = \{h_{ij}\}$:

$$h_{ij} = \begin{cases} 0, & Q_{b_o} \cap Q_{z_{ij}} = \varnothing \\ 1, & Q_{b_o} \cap Q_{z_{ij}} \neq \varnothing \end{cases} \quad , \quad i, j = 0, ..., 4 \tag{12}$$

where h_{ij} represents the existence of intersection of b_o with z_{ij}. This encoding method can not only describe all possible spatial relationships in a consistent way but also make a distinction between $\langle$*subject, object*$\rangle$ and $\langle$*object, subject*$\rangle$. Then, H is reshaped to a vector $s \in \mathbb{R}^{25}$. The spatial feature is extracted from s with a fully connected layer and then fuse with the conventional encoding feature.

3.3 Scene Graph Generation

A scene graph consists of the class labels with the locations of individual objects and the relationship between each pairwise object [9], which can be defined as :

$$G = \{B, O, R\} \tag{13}$$

where $B = \{b_1, b_2, ..., b_n\}$ is a set of bounding boxes, $O = \{O_1, O_2, ..., O_n\}$ is the set of class labels corresponding to B, which is refined in Sec. 3.1 with the self-study method, $R = \{r_{O_0 \to O_1}, r_{O_0 \to O_2}, ..., r_{O_n \to O_{n-1}}\}$ is the set of relation between O_i and O_j with $n(n-1)$ elements. The relationships R is predicted with a Roi Relation Head. In this paper, we use MOTIFS [2], as the Roi Relation Head, and debias the predict of the relation R with TDE [20]. Finally, the triplet list is ranked with the comprehensive confidence score of the object and the predicate. The scene graph is generated with the combination of the detected pairwise objects and their predicted relationships, and finally ordered by its joint probability $P(O_i)P(R_{O_i \to O_j})P(O_j)$.

4 Experiment

4.1 Experimental Settings

Datasets. Following the recent works [9,20,2] in SGG, we trained and evaluated our model on the Visual Genome (VG) [26] dataset. It consists of 108k images with 75k object categories and 37k predicate classes. Since 92% predicate classes have no more than 10 samples, we followed previous works [1] and adopted a widely used VG split, containing the 150 most frequent object categories with 50 predicate classes. Meanwhile, the VG dataset is split into a training set (70%) and a test set (30%) with a validation set (5k) sampled from the training set for parameter tuning.

Table 1. The SGG performances of Relationship Retrieval on mean Recall@K [9,23], and the CSS is our proposed model.

		Predicate Classification			Scene Graph Classification			Scene Graph Detection		
Model	Method	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
IMP+ [1,9]	-	-	9.8	10.5	-	5.8	6.0	-	3.8	4.8
FREQ [23,2]	-	8.3	13.0	16.0	5.1	7.2	8.5	4.5	6.1	7.1
KERN [9]	-	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
PA [61]	-	15.2	19.2	20.9	8.7	10.9	11.6	5.7	7.7	8.8
GPS-Net [62]	-	17.4	21.3	22.8	10.0	11.8	12.6	6.9	8.7	9.8
GB-Net-β [63]	-	-	22.1	24.0	-	12.7	13.4	-	7.1	8.5
VTranseE [24]	baseline	11.6	14.7	15.8	6.7	8.2	8.7	3.7	5.0	6.0
	TDE [24]	18.9	25.3	28.4	9.8	13.1	14.7	6.0	8.2	10.2
MOTIFS [2]	baseline	10.8	14.0	15.3	6.3	7.7	8.2	4.2	5.7	6.6
	Focal	10.9	13.9	15.0	6.3	7.7	8.3	3.9	5.3	6.6
	Reweight	16.0	20.0	21.9	8.4	10.1	10.9	6.5	8.4	9.8
	Resample	14.7	18.5	20.0	9.1	11.0	11.8	5.9	8.2	9.7
	Lu+cKD [64]	14.4	18.5	20.2	8.7	10.7	11.4	5.8	8.1	9.6
	CogTree [22]	20.9	26.4	29.0	12.1	14.9	16.1	7.9	10.4	11.8
	TDE [20]	18.5	24.9	28.3	11.1	13.9	15.2	5.8	8.2	9.8
	TDE-CSS	**20.0**	**26.1**	**28.5**	**11.8**	**14.8**	**16.2**	**6.7**	**8.9**	**10.8**
	DLFE [25]	22.1	26.9	28.8	12.8	15.2	15.9	8.6	11.7	13.8
	DLFE-CSS	**23.9**	**28.8**	**30.7**	**13.6**	**16.0**	**16.9**	**8.7**	**12.0**	**14.1**
VCTree [23]	baseline	14.0	17.9	19.4	8.2	10.1	10.8	5.2	6.9	8.0
	Reweight	16.3	19.4	20.4	10.6	12.5	13.1	6.6	8.7	10.1
	Lu+cKD [64]	14.4	18.4	20.0	9.7	12.4	13.1	5.7	7.7	9.1
	CogTree [22]	22.0	27.6	29.7	15.4	18.8	19.9	7.8	10.4	12.1
	EBM [65]	14.2	18.2	19.8	10.4	12.5	13.5	5.7	7.7	9.1
	EBM-CSS	**17.1**	**20.8**	**22.3**	**10.9**	**13.0**	**14.1**	**6.0**	**7.1**	**9.7**
	TDE [20]	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
	TDE-CSS	**19.4**	**25.9**	**29.4**	**9.2**	**12.9**	**14.9**	**7.1**	**9.6**	**11.8**
	DLFE [25]	20.8	25.3	27.1	15.8	18.9	20.0	8.6	11.8	13.8
	DLFE-CSS	**23.7**	**28.6**	**30.5**	**16.0**	**18.9**	**20.4**	**8.7**	**11.9**	**14.0**

Task and Evaluation. We followed the previous work [2] to divide the SGG task into three sub-tasks: (1) Predicate Classification (**PredCls**) which takes the ground truth bounding boxes with its object labels for relation prediction; (2) Scene Graph Classification (**SGCls**) which takes ground-truth bounding boxes to predict the object label and the relation between the pairwise objects. (3) Scene Graph Detection (**SGDet**) which detects scene graph from scratch. The metric of the traditional SGG task is **Recall@K(R@K)**, which is the fraction of ground-truth targets that are recalled correctly in top K predictions [12]. Due to the long-tailed bias, the good performance on R@K caters to "head" predicates, e.g. *on* [20]. The metric of the recent unbiased SGG task is **mean Recall@K(mR@K)** [9,23], which retrieves each class of relation separately and averages R@K for each relation. The good performance on mR@K achieves more balanced results among different predicates.

Model Configuration. In this paper, we evaluated our method with the roi relation head based on two classic baselines: MotifNet [2] and VCTree [23]. The

Table 2. The results of Relationship Retrieval on Recall@K, , and the CSS is our proposed model. The Motifs-TDE and VCTree-TDE are traditional SGG approaches, and the others are the unbiased SGG approaches.

	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
Model	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
IMP+ [1,9]	52.7	59.3	61.3	31.7	34.6	35.4	14.6	20.7	24.5
FREQ [23,2]	53.6	60.6	62.2	29.3	32.3	32.9	20.1	26.2	30.1
KERN [9]	-	65.8	67.6	-	36.7	37.4	-	27.1	29.8
VTransE [24]	59.0	65.7	67.6	35.4	38.6	39.4	23.0	29.7	34.3
Motifs-TDE [20,2]	38.7	50.8	55.8	21.8	27.2	29.5	12.4	16.9	20.3
VCTree-TDE [20,23]	39.1	49.9	54.5	22.8	28.8	31.2	14.3	19.6	23.3
MOTIFS [2]	58.5	65.2	67.1	32.9	35.8	36.5	21.4	27.2	30.3
MOTIFS-CSS	**59.5**	**66.1**	**67.9**	**35.9**	**39.1**	**39.9**	**25.2**	**32.3**	**37.2**
VCTree [23]	60.1	66.4	68.1	35.2	38.1	38.8	22.0	27.9	31.3
VCTree-CSS	**61.6**	**66.9**	**68.5**	**41.6**	**45.6**	**46.6**	**24.5**	**31.4**	**36.0**

Table 3. Ablation studies of individual components of our method. The baseline model mentioned below is Motifs-TDE unless otherwise indicated.

		Predicate Classification		Scene Graph Classification		Scene Graph Detection	
SS	SA	mR@20/R@20	mR@50/R@50	mR@20/R@20	mR@50/R@50	mR@20/R@20	mR@50/R@50
-	-	18.5 / 38.7	24.9 / 50.8	11.1 / 22.1	13.9 / 27.2	5.8 / 12.4	8.2 / 16.9
✓	-	- / -	- / -	11.4 / 21.2	14.6 / 27.9	6.7 / 12.9	8.9 / 17.5
-	✓	**20.0** / **42.0**	**26.1** / **53.1**	11.2 / 24.8	14.5 / 30.4	6.4 / 13.3	8.7 / 18.4
✓	✓	- / -	- / -	**11.8** / **26.2**	**14.8** / **31.7**	**6.4** / **12.9**	**8.9** / **18.6**

fusion function for the relation head is set to sum in PredCls and SGDet, and gate in SGCls. Other hyperparameters can be viewed in **Model Zoo** [66]. All models share the same pre-trained detector and the same settings as well.

4.2 Implementation Details

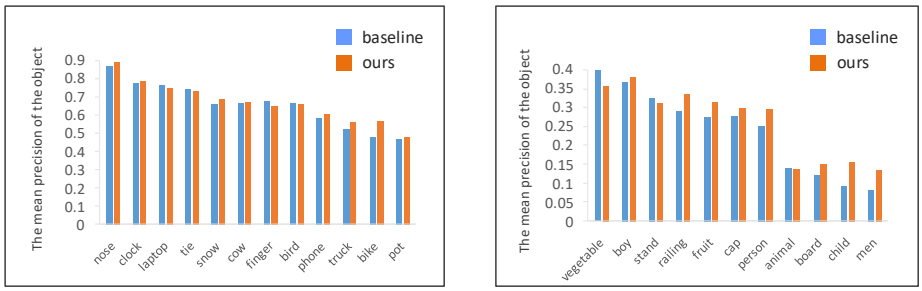
Following the previous work [20], we used a pre-trained Faster R-CNN [58] with a ResNeXt-101-FPN [67,68] and freeze the weights during the training process. For SGCls and SGDet tasks, we first train the original SGG models with the source domain recommended from the configuration for all tasks, including the learning rate. The batch size is set to 12. Then we initialized the Memorandum with the statistics of the results for its counterparts, respectively. The attenuation factor α is set to 0.998 in this paper.

4.3 Experiment Results

We evaluated the CSS with the comparison of the conventional unbiased SGG approaches and traditional SGG approaches. As illustrated in Tabs. 1 and 2, we compared our performance with several state-of-the-art unbiased SGG methods:

Table 4. The average precision (left) and the average recall (right) of the object detection for the bounding boxes with different sizes.

	Predicate Classification				Scene Graph Detection			
SS		✓		✓		✓		✓
SA			✓	✓			✓	✓
small	46.2/53.6	**47.4**/**54.7**	46.4/53.7	47.2/**54.7**	5.7/21.7	5.7/**21.8**	**5.8**/21.7	**5.8**/**21.8**
medium	54.8/62.0	**55.6**/62.8	55.1/62.0	55.6/**62.9**	11.9/32.3	11.9/32.3	11.9/32.3	11.9/32.3
large	53.1/60.7	53.9/61.4	53.2/60.7	**54.0**/**61.5**	17.9/35.4	17.9/**35.6**	17.9/35.4	17.9/**35.6**
all	56.6/64.2	**57.4**/**64.9**	56.8/64.2	**57.4**/**64.9**	12.9/34.4	**13.0**/**34.5**	12.9/34.4	**13.0**/**34.5**



(a) The average precision of the easy sample

(b) The average precision of the hard sample

Fig. 5. The mean precision of the object between the easy and hard sample in the SGCls task. The results of objects with high precision are close between the baseline and CSS, while CSS obtain an obvious improvement to the low-precision objects.

TDE [20], EBM [65] and DLFE [25] with mR@K, and the classic traditional SGG model: Motifs [2] and Vctree [23] with R@K.

Object Retrieval. We accumulated the average precision and the average recall of objects with different sizes by the COCO-API [69]. As illustrated in Tab. 4, both the precision and the recall achieve promotion from the baseline with an average of 1.4% and 1.5% relative gain in SGCls sub-task. However, the optimization of both the precision and recall are under 0.5% in SGDet.

Relationship Retrieval (RR). The results are listed in Tabs. 1 and 2. The CSS model improves on the baseline by an average of 4.6%, 4.5%, 9.3% relative gain of mR@K in each subtask respectively. Meanwhile, it is obvious that the debiasing method causes an unacceptable decline in Recall, as shown with the unbiased SGG model in Tab. 2. Moreover, the recall of each predicate is applied in the supplementary materials. The CSS improves the RR with an average of 37.8% of the head predicates and 26.72% of the tail predicates, which can be illustrated intuitively in Fig. 6.

Ablation Study. We considered the ablations of each module to investigate the effectiveness of each part of the proposed CSS. The results of SS and CSS are vacant in PredCls task because the label of the object has already been provided as the input of this task, therefore there is no need for object refinement.

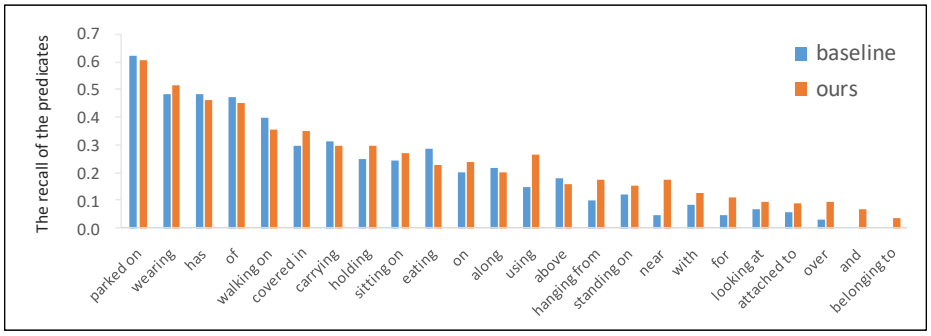


Fig. 6. The recall of the relationships between baseline and CSS. It indicated that the recall of the overall predicates is improved with CSS.

As illustrated in Table. 3, each module obtains an obvious promotion: over the baseline, the SS and SA module achieves an average of 8.2% and 5.2% respectively. Besides, we observed that the average precision and recall between SS and CSS are close in Table. 4 which obtain an obvious improvement, while the results of SA are almost consistent with the baseline. Moreover, the recall of each predicate shows that the SS module improves the tail predicates with an average of 40% without any optimization of the head predicates, while SA promotes the RR with 19.3% of the head predicates and 32.6% of the hard ones. In addition, we evaluated the influence of the granularity of the relative position encoding in the scene augmentation module. The granularity is set to be 5×5, 8×8 and 13×13. The results are recorded in the supplementary materials, which show limited promotion with the increase of the encoding granularity. However, the increased cost of time and resources is unacceptable.

4.4 Quantitative Studies

Object Detection. As illustrated in Tab. 4, the SS module achieves promising improvement on object detection in SGCls, especially the targets with small bounding boxes. However, the promotion is limited in the SGDet sub-task. Since the scene information of complex image contains unexpected noise, the SS module is struggle to refine objects with confounding factors. While the labeling process selects the bounding box with human focus, it naturally mitigates the scene noise in SGCls. With the denoising of the bounding box proposals, the SS has great potential to optimize the objects in complex scenes in the future.

Visual Relation Detection. The RR results verify that CSS can refine the SGG effectively with the promising promotion. The recall among each relation shows that SA optimizes the prediction of the overall predicates, while SS focuses more on the tail predicates. Combining the results of SS in Tabs. 3 and 4, it shows that the prediction of the relationships is sensitive to the object, which enables the SS module still work with limited refinement of the object detection. Further, we can infer that the hard sample has a strong constraint on the tail predicates.

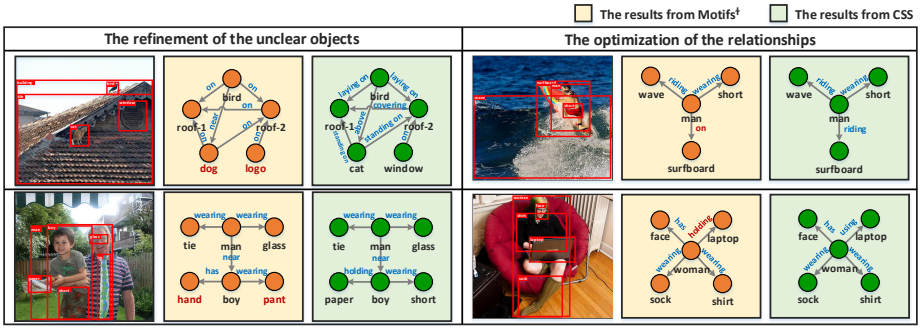


Fig. 7. The visualization results of SG generated from MOTIFS-TDE [20] and CSS.

Scene Graph Generation. The improvement of both the R@K and mR@K illustrated the effectiveness of the CSS model. Moreover, we observed that the CSS achieves the best performance in mR@K and R@K from different aspects. It illustrates that the two modules have different preferences and the CSS balances the results of the head and the tail predicates.

4.5 Visualization Results

For a more intuitive explanation, we generated several SGCls examples from MOTIFS-TDE and Motifs-CSS. As illustrated in Fig. 7, the first row shows the optimization of the unclear objects with small bounding boxes. The top example of the first row also consistent with our analysis in Sec. 4.4 that the relationships are sensitive with the objects. With the refinement of *cat* from *dog*, the misclassified predicate *on* is also optimized with *standing on* correctly. The second row shows examples of the debiasing work. We can observe the refinement of both the head predicate *riding* and the tail predicate *using*, which can provide richer information for the downstream tasks.

5 Conclusion

In this work, we introduced a Continuous Self-Study model (CSS) for scene graph generation. By learning the self-behavior and combining the scene information, the CSS improves the accuracy of identifying the ambiguous targets in complex images. Meanwhile, with the fine-grained relative position encoding, the CSS is able to discriminate the visual relationships with long-tailed bias effectively. Since our proposed method achieves improvements of two basic tasks: object detection and visual relationship detection, it will be helpful to improve the performance in much more visual understanding tasks in the future.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (NSFC No.62076031).

References

1. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 5410–5419
2. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5831–5840
3. Ye, K., Kovashka, A.: Linguistic structures as weak supervision for visual scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 8289–8299
4. Wang, Y.S., Liu, C., Zeng, X., Yuille, A.: Scene graph parsing as dependency parsing. arXiv preprint arXiv:1803.09189 (2018)
5. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification. arXiv preprint arXiv:1612.04844 (2016)
6. Wang, W., Wang, R., Shan, S., Chen, X.: Exploring context and visual pattern of relationship for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 8188–8197
7. Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., Wang, X.: Factorizable net: an efficient subgraph-based framework for scene graph generation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 335–351
8. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 1969–1978
9. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6163–6171
10. Hung, Z.S., Mallya, A., Lazebnik, S.: Contextual translation embedding for visual relationship detection and scene graph generation. IEEE transactions on pattern analysis and machine intelligence (2020)
11. Fang, Y., Kuan, K., Lin, J., Tan, C., Chandrasekhar, V.: Object detection meets knowledge graphs, International Joint Conferences on Artificial Intelligence (2017)
12. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE conference on computer vision and Pattern recognition. (2017) 3076–3086
13. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE international conference on computer vision. (2017) 1261–1270
14. Chiou, M.J., Liu, Z., Yin, Y., Liu, A.A., Zimmermann, R.: Zero-shot multi-view indoor localization via graph location networks. In: Proceedings of the 28th ACM International Conference on Multimedia. (2020) 3431–3440
15. Armeni, I., He, Z.Y., Gwak, J., Zamir, A.R., Fischer, M., Malik, J., Savarese, S.: 3d scene graph: A structure for unified semantics, 3d space, and camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5664–5673
16. Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D.: Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In: Proceedings of the fourth workshop on vision and language. (2015) 70–80
17. Norcliffe-Brown, W., Vafeias, E., Parisot, S.: Learning conditioned graph structures for interpretable visual question answering. arXiv preprint arXiv:1806.07243 (2018)

18. Zhu, Z., Yu, J., Wang, Y., Sun, Y., Hu, Y., Wu, Q.: Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering. arXiv preprint arXiv:2006.09073 (2020)
19. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 10685–10694
20. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 3716–3725
21. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 16383–16392
22. Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q.: Cogtree: Cognition tree loss for unbiased scene graph generation. arXiv preprint arXiv:2009.07526 (2020)
23. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 6619–6628
24. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 5532–5540
25. Chiou, M.J., Ding, H., Yan, H., Wang, C., Zimmermann, R., Feng, J.: Recovering the unbiased scene graphs from the biased ones. In: Proceedings of the 29th ACM International Conference on Multimedia. (2021) 1581–1590
26. 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. arXiv preprint arXiv:1602.07332 (2016)
27. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). (2018) 670–685
28. Ji, M., Shin, S., Hwang, S., Park, G., Moon, I.C.: Refine myself by teaching myself: Feature refinement via self-knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 10664–10673
29. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
30. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 3713–3722
31. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5007–5016
32. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, PMLR (2021) 8748–8763

34. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision, Springer (2016) 852–869
35. Shi, J., Zhang, H., Li, J.: Explainable and explicit visual reasoning over scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 8376–8384
36. Krishna, R., Chami, I., Bernstein, M., Fei-Fei, L.: Referring relationships. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6867–6876
37. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 1219–1228
38. Woo, S., Kim, D., Cho, D., Kweon, I.S.: Linknet: Relational embedding for scene graph. arXiv preprint arXiv:1811.06410 (2018)
39. Dunn, G., Emsley, R., Liu, H., Landau, S., Green, J., White, I., Pickles, A.: Evaluation and validation of social and psychological markers in randomised trials of complex interventions in mental health: a methodological research programme. Health Technology Assessment (Winchester, England) **19** (2015) 1–116
40. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 3957–3966
41. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
42. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
43. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
44. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. arXiv preprint arXiv:1802.04977 (2018)
45. Ahn, S., Hu, S.X., Damianou, A., Lawrence, N.D., Dai, Z.: Variational information distillation for knowledge transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 9163–9171
46. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4133–4141
47. Koratana, A., Kang, D., Bailis, P., Zaharia, M.: Lit: Learned intermediate representation training for model compression. In: International Conference on Machine Learning, PMLR (2019) 3509–3518
48. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 7096–7104
49. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 3967–3976
50. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE (2020) 1–8
51. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 11557–11568

52. Xu, T.B., Liu, C.L.: Data-distortion guided self-distillation for deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 5565–5572
53. Yun, S., Park, J., Lee, K., Shin, J.: Regularizing class-wise predictions via self-knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 13876–13885
54. Lee, H., Hwang, S.J., Shin, J.: Self-supervised label augmentation via input transformations. In: International Conference on Machine Learning, PMLR (2020) 5714–5724
55. Lan, X., Zhu, X., Gong, S.: Knowledge distillation by on-the-fly native ensemble. arXiv preprint arXiv:1806.04606 (2018)
56. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
57. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 6848–6856
58. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 91–99
59. Vollmer, M.: Newton's law of cooling revisited. European Journal of Physics **30** (2009) 1063
60. Yang, Z., Huang, X., Xiu, J., Liu, C.: Socialrank: Social network influence ranking method. In: 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems. Volume 2., IEEE (2012) 591–595
61. Tian, H., Xu, N., Liu, A.A., Zhang, Y.: Part-aware interactive learning for scene graph generation. In: Proceedings of the 28th ACM International Conference on Multimedia. (2020) 3155–3163
62. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 3746–3753
63. Zareian, A., Karaman, S., Chang, S.F.: Bridging knowledge graphs to generate scene graphs. In: European conference on computer vision, Springer (2020) 606–623
64. Wang, T.J.J., Pehlivan, S., Laaksonen, J.: Tackling the unannotated: Scene graph generation with bias-reduced models. arXiv preprint arXiv:2008.07832 (2020)
65. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.: Energy-based learning for scene graph generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2021) 13936–13945
66. Tang, K.: A scene graph generation codebase in pytorch (2020) `https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch`.
67. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2117–2125
68. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 1492–1500
69. 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: European conference on computer vision, Springer (2014) 740–755