Energy-Based Learning in Weakly Supervised Scene Graph Generation

Berkay Güler
KU Leuven
Leuven, Belgium
berkay.guler1@gmail.com

***Abstract*—Nowadays, the literature about scene graph generation (SGG) focuses more on weak supervision in training. That is mainly caused by the fact that full supervision possesses demanding human effort, and inherent bias and errors due to the annotators. The weakly supervised setting overcomes these issues since there is no need for annotations for the images, thus making it also more scalable and cheaper. Besides the problems in training scheme, the traditional approaches for the SGG task utilize a sum of cross-entropy losses for the detected object pairs ignoring the composition present in the image. This thesis aims to utilize a proposed energy-based framework to understand the structure in the image, and to improve the results of a baseline relation model under weak supervision. To achieve this goal, a state-of-the-art relation model is modified to make it compatible with weakly supervised setting. Similar modifications are applied to the energy model and these two models are integrated into each other. According to the experimental results, the modified energy framework[1] enhances the performance of the baseline model for the rare predicate words marginally, and it follows the same improvement trend for these rare relation categories even under weak supervision.**

***Index Terms*—SGG, Energy Learning, Weak Supervision**

I. Introduction

As the area of computer vision grows each year, traditional tasks such as object detection are not considered to be exciting anymore. Therefore, the researchers initiated more instructive challenges like scene graph generation (SGG) to get a better representation of the images. This graph-based representation is also beneficial to other several related applications, including visual question answering [1], image captioning [2], and image retrieval [3].

A typical scene graph model tries to recognize the visual semantics in the image and detects the triplets of $\langle subject, predicate, object \rangle$. The detector identifies the locations and categories of objects, and the overall model finds the predicate word meaning the relationship between this pair. The detection of these triplets could be done in two main ways namely one-staged or two-staged methods. This thesis mainly focuses on the two-staged methods. In this method, the proposal regions and features are extracted by an object detector. Then, these proposals are utilized to find the relation category between the proposals.

The training scheme of a typical SGG model primarily considers two ways. The first one is full supervision, where the bounding box annotations are present the in the ground truth. For the classification stage, these box annotations provide detailed information about the relationships between each object pair. In contrast, weak supervision does not possess any bounding box annotations. The literature for the weak supervision [4]–[9] utilizes different weak supervision schemes. The most popular ones are the image-level relation labels and the ground truth image-level graphs. These popular schemes are easier cases than employing image captions and are mainly used to provide an upper limit before applying image caption training. Although SGG is initially designed for full supervision, weak supervision studies started to gain attention in recent years. This is mainly caused by the fact that full supervision annotations are expensive and prone to annotator errors and biases. Also, caption-based supervision is easier to collect making the weak supervision more scalable. Therefore,

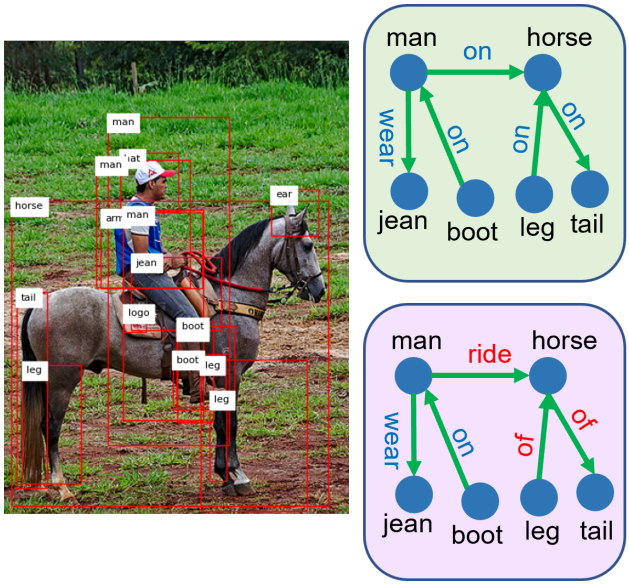


Fig. 1: **Scene Graph Generation:** Figure illustrates an example scene graph generation both from the baseline cross-entropy model (green) and energy model (purple). It is clearly observed that the baseline model is biased towards to the common predicates like `on, wear`; however, the energy model incorporated the composition of the image and identified the rare predicates `ride, of`.

[1]Code: https://github.com/gulerrberkay/scene-graph-ebm

this paper mostly focuses on weak supervision thanks to these benefits.

The weakly supervised SGG contains fundamental challenges that have to be confronted. Firstly, such weak supervision does not provide as detailed information as full supervision during training. Secondly, most of the studies in weak supervision benefits from an off-the-shelves object detector to identify the proposed regions for the object. However, these detectors usually produce some noisy object predictions along with confident ones. One has to trust these predictions in the training phase to train their model. Thirdly, a typical two-stage scene graph structure aggregates the context information using several methods. These models are usually trained with the sum of cross-entropy losses in weak and full supervision cases. However, applying such a loss technique lacks information on the composition of the image, and it treats every object pair independently. By commonsense, it is known that most objects in the image are somehow related, so this underlying structure has to be considered in the loss function. Fourthly, the well-known dataset Visual Genome [10] has a long-tailed distribution for the predicate words, so the problem of SGG has a biased training process. This bias causes the conventional models to predict the frequent predicate more than the rare ones.

In recent years, energy-based learning strategies start to get promising results for the image generation studies like [11], [12]. Also, they have an increasing usage in discriminative studies such as [13], [14]. These energy frameworks assign scalar energy values for the input-output configurations. Given the input x with label y, the joint energy model can be denoted as $E_\theta(x, y)$. [15] states that any function could be selected if one satisfies the two main criteria of probability distributions: **i)** the probability distribution needs to assign a non-negative value for every input value x i.e. $p(x) \geq 1$. **ii)** the integral of probability distribution needs to be equal to 1 for each input i.e. $\int_a^b x^2\, dx = 1$. This could be achieved by a Boltzmann distribution as in $q_\theta(x) = \frac{\exp(-E_\theta(x))}{Z_\theta}$. [15] also mentions that the denominator of this crafted distribution is usually untraceable. Thus, it is not trivial to find the best parameters that maximize the likelihood. According to [15], most methods address this issue by writing the derivate of log-likelihood as $\nabla \mathcal{L}_{MLE}(\theta; p) = \mathbb{E}_{p(x)}[\nabla_\theta E_\theta(x)] - \mathbb{E}_{q_\theta(x)}[\nabla_\theta E_\theta(x)]$. This new representation requires Monte Carlo Markov Chain (MCMC) methods to sample from the data distributions to compute the expectation.

This paper proposes using energy-based frameworks in weak supervision to address all aforementioned problems. It has been proved that the energy-based approach improves baseline cross-entropy models under full supervision thanks to the study of [13]. Therefore, exploiting such energy formulation should guide the learning process of the baseline model leading the better results. According to [13], the main reason behind the success of the energy models is that they convert the problem of maximization of the sum of likelihoods into the maximization of joint likelihood problem by taking the structure in the input image into account. Thus, the main research question in this paper is how much the energy models improve the baseline relation model results even under weak supervision.

Contributions. The summary of contributions are as follows in this paper:

- A conventional relation model is adjusted and made compatible with the current weak supervision setup. The baseline results established solid fundamentals for the next part.
- The energy framework is also modified for the weakly training. The integration of the energy model into the relation model is done to see any improvements in the results for the weak supervision.
- To realize the stated integration, a filtering and sampling mechanisms under weak supervision are proposed. The ablation studies for each additional method are provided to illustrate their effectiveness.

II. Related Works

Weakly Supervised SGG: Scene graph generation is one of the popular tasks in vision-related research, and in recent years various papers have been shared mostly under full supervision. The studies like [1], [16]–[20] utilizes the full supervision training scheme to tackle the SGG task. However, as mentioned earlier weak supervision gained popularity against full supervision because of the inherent problems of fully supervised training. The weakly supervised SGG papers like [4]–[9] utilize similar strategies while training their model. VSPNet [4] utilizes the image-level relation labels, and the ground truth image graph as the ground truth to align their predictions with this ground truth graph. The study of [5] extends to problems from ground truth graphs to image captions to make the problem setup even weaker. Their method wants to benefit from the linguistic structure present the captions to assist the training under weak supervision. [6] employs a simpler approach to find the relations between the objects. Their method contains a sensitivity analysis that detects the effect of each object on the predicted relations to find out which objects produced the high-scored relations. Both studies of [8], [9] consider turning the weak supervision problem into a fully supervised one by generating pseudo ground truths graphs. [8] achieves the goal by applying graph alignment algorithms on the detected proposals whereas [9] utilizes a previously trained grounding module to localize each detected proposal. These localized graphs become the pseudo-ground truths, transforming the problem into a fully supervised one.

Energy Based Modeling: In recent years, the energy-based methods started to gain attention in image generative studies like [11], [12]. These papers exploit the energy model architecture to increase the image generation accuracies. Even though the energy models are often popular in generative tasks, the studies [13], [14] focusing on discriminative tasks also started to use the energy models to get improvements in their results. Since [13] proved that the energy framework is highly

useful for the SGG task, applying this methodology in the weakly supervised SGG training seems to be promising.

III. Method

This section introduces the proposed methodology in the paper. Section III-A introduces the main variables in the proposed method. Section III-B discusses each layer in the architecture along with the formulations of the structure. Section III-C investigates the overall loss functions for the training of the designed architecture.

A. Problem Setup

This subsection introduces the proposed problem setup for weakly supervised SGG. Given image I, the scene graph model $\mathcal{M}$ calculates a tuple of (O, R) i.e. $\mathcal{M}(I) = (O, R)$ where $O \in \mathbb{R}^{NxN_o}$ represents the detected objects in image I. The relations between the detected objects are shown in $R \in \mathbb{R}^{(N*N-1)xN_r}$. The number of proposals, object classes, and relation classes are notated as N, N_o, N_r respectively in the previous formulas.

For weak supervision, only image-level ground truths are available for the baseline model. That means for a given image I, the ground truth object labels for this image $\mathcal{O}_I$ is a subset of $\mathcal{O} \in \{0, 1, ...N_o\}$. For the same image, image-level ground truth relation labels $\mathcal{R}_I$ is a subset of $\mathcal{R} \in \{0, 1, ...N_r\}$. In addition, the dimension of the relation score matrix R increases quadratically by the number of proposals in weak supervision because it needs to calculate the relation scores for all pairs.

B. Model Design

This subsection investigates the model overview depicted in Figure 2 layer by layer and provides how the adjusted Motif [16] and energy framework [13] functions.

1) Relation Model: This paper chooses the architecture of Motif [16] which is considered one of the traditional LSTM-based relation models. It utilizes the natural ability of Recurrent Neural Network (RNN) based architectures to detect the relationships between the objects. Although using proposals as a sequence input for the LSTMs seems counter-intuitive, the LSTM structures are highly utilized in the SGG community due to their success. The contribution that converts Motif suitable in weakly supervised training is the elimination of the Decoder part in the original Motif [16]. This layer was decoding the proposals and assigns refine object labels for them. However, it cannot be utilized in weak supervision since the ground truth labels for the proposal boxes are not available.

The relation model partitions the problem of relation detection with equation 1.

$$Pr(G_{SG}|I) = Pr(B|I)Pr(O|B, I)Pr(R|B, O, I) \quad (1)$$

The first and second terms $Pr(B|I), Pr(O|B, I)$ in equation 1 denotes the problem of object detection. An off-the-shelf object detector detects proposal regions for the input images and provides the predicted object distributions. The labels of these proposals can be calculated from the maximum-scored object category. The last term $Pr(R|B, O, I)$ illustrates the relation identification part where the detected object labels, and locations of these objects are all being used to find the scene graph.

The relation model first extracts the features and labels thanks to the detector as mentioned earlier. The predicted bounding boxes $B = \{b_1, ..., b_N\}$, feature vectors $f_i \in \mathbb{R}^{4096}$ and labels $l_i \in \mathbb{R}^{200}$ are detected for the i-th proposal.[2]. [16] argues that these detected proposals need to be contextualized by using them in biLSTM layers. Hence, the next step is called the *object context* where these objects produce a better representation of the given image I.

Object Context. A sequence of inputs is created by the proposals by feeding them into the first biLSTM layer called object context. The default option from [16] is selected for ordering the proposals. This default option orders the proposals from left to right by looking at the detected bounding box.

$$C = biLSTM\left([f_i\,;\mathbf{W_1}\mathbf{l}_i]_{i=1,...,n}\right) \quad (2)$$

Equation 2 provides the formulation of the first layer where $C = [c_1, ..., c_N]$ are hidden states for each proposal. $\mathbf{W_1}$ is a mapping kernel for arranging the dimension of l_i.

Edge Context. The next step is to compute a relation matrix $R \in \mathbb{R}^{(N*N-1)xN_r}$. This matrix holds all relation scores for every object pair. That is why its dimension rises rapidly if one chooses to increase the number of proposals. The contextualized proposals in C serve as inputs in this layer to apply a similar procedure on edge vectors.

$$D = biLSTM\left([\mathbf{c}_i\,;\mathbf{W_2}\mathbf{l}_i]_{i=1,...,n}\right) \quad (3)$$

$D = [d_1, ..., d_n]$ represents the edge vectors for the proposals. However, these representation lacks global knowledge about the object pair. To incorporate this additional information, the union's visual features need to be mixed into the solution. Moreover, most of the studies [1], [6], [16], [17], [19] mentioned in Section II activates the frequency baseline which puts more bias towards some relations if a particular object pair is detected. After adding this information, equation 4 displays how to compute the matrix R for relation scores.

$$R = (\mathbf{W_h}d_i \circ \mathbf{W_t}d_j) \circ f_{i,j} + w_{i,j} \in \mathbb{R}^{(N*N-1)xN_r} \quad (4)$$

$\mathbf{W_2}, \mathbf{W_h}, \mathbf{W_t}$ are dimension mapping matrices for d_i. R matrix holds every relation score for each object pair; thus, this graph representation can be used in the energy model.

Inference. The evaluation of the relation model needs sorting operation on the triplets. Each predicted triplet has three components $\langle subject, predicate, object\rangle$. The scores of each component are calculated by performing a softmax operation on matrices R, O. Three scores will be multiplied to obtain the score for the corresponding triplet. The most confident triplets will be considered in the evaluation part.

[2]Note that l_i has a dimension of $\mathbb{R}^{200}$ since the loaded GloVe embeddings [21] has a vector of this dimension.

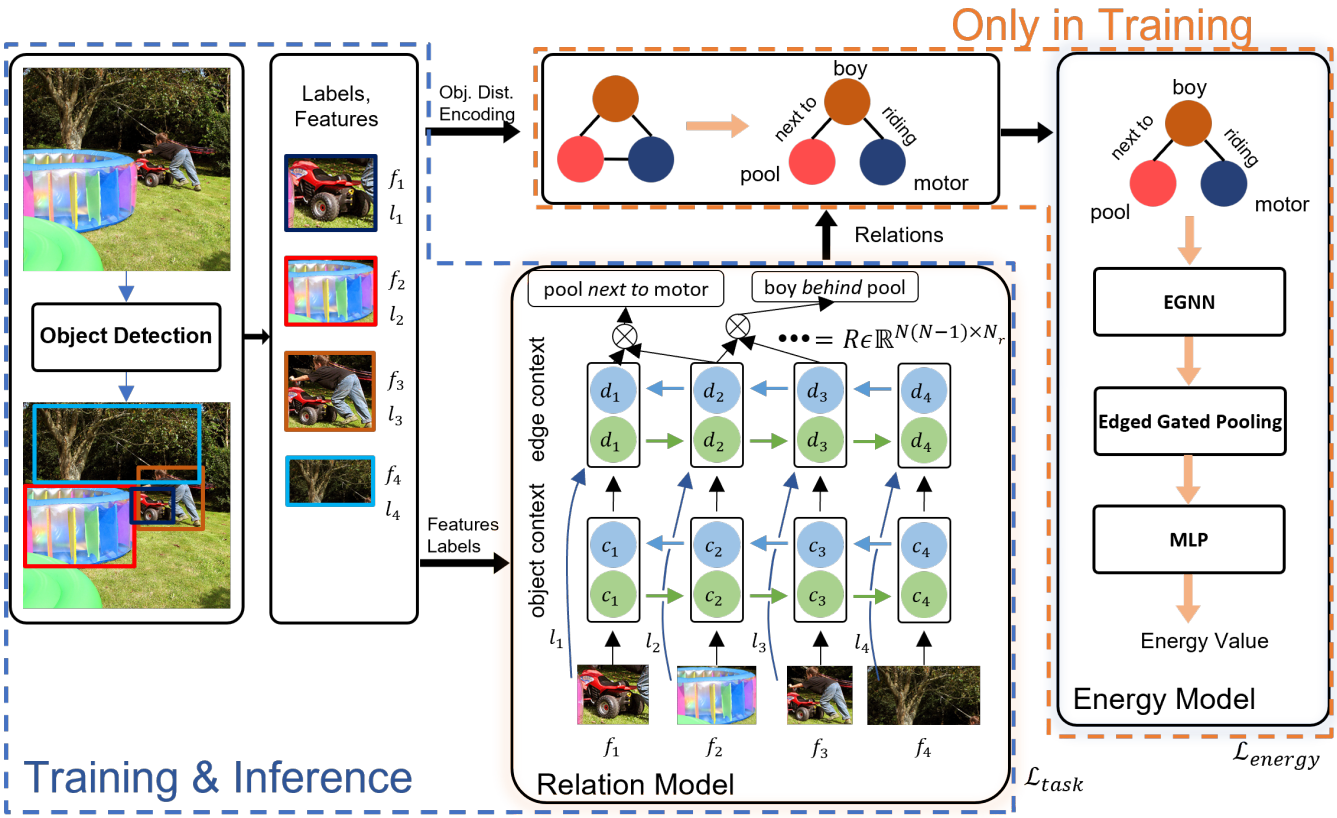


Fig. 2: **Model Overview.** The overall pipeline starts from the object detector on the left to find the labels, features, the bounding boxes. These features and labels are fed into the relation model to make them contextualized. The edge context utilizes the inputs from labels and object context to calculate the relation score matrix R. The relation model is always activated both for training and inference. The energy model, on the other hand, is only activated in training. It guides the training procedure for the relation model by calculating the energy loss from the predicted and ground truth scene graphs.

2) Energy Model: This subsection analyzes the energy model architecture depicted in Figure 2. An important remark about the energy model is that it is only activated in the training phase. Thus, it is not employed in the inference phase. The most important contribution is the removal of the image graph from the original work [13]. The image graph contains the RoI features from the proposals and ground truth object boxes; hence, it incorporates the visual features in the energy loss. Since it is not suitable for weak supervision due to the absence of box annotations, this image graph is removed from the formulations. The effects of this removal can be seen in Section IV.

Overview. The energy model in Figure 2 starts from the encoded scene graph predicted by the relation model. Therefore, the problem setup for the energy model can be summarized as follows: Given a scene graph generation model $\mathcal{M}$, and an image I, the predicted scene graph can be denoted as $G_{SG} = \mathcal{M}(I) = (R, O)$. The tuple of (O, R) is two main inputs of the energy model as they represent the scene graph. This tuple undergoes Edged Gated Neural Network (EGNN) and pooling layers proposed by the authors of [13]. These layers essentially apply message passing between the detected objects and relations to get a meaningful representation. The pooling functions decrease the dimension of the hidden embeddings into one scalar value called the energy.

$$E_\theta(G_{SG}) = MLP[f(EGNN(G_{SG}))] = e^- \quad (5)$$

$$E_\theta(G^+_{SG}) = MLP[f(EGNN(G^+_{SG}))] = e^+ \quad (6)$$

Equations 5 and 6 denote the general formulation of the energy model. e^- is called the negative energy and holds the information for the predicted graph configuration. e^+ is referred to as the positive energy and it is the scalar representation of the ground truth configuration. Therefore, **the energy model needs the ground truth image-level graph to calculate the energy loss.** The image-level graph should guide the training by considering the structural information present in the image. The graph-level representation in the ground truth is still considered weak supervision as it does not utilize box annotations. For instance, an image caption can be turned into an image-level graph by a language parser [22].

EGNN Layer. According to [13], this layer aims to enrich each node and edge state by applying message passing algorithms between them. One should note that in energy model,

the nodes represent object proposal, and the edges represent the relations between the object pairs.

These nodes and edges are predicted from the modified Motif structure; thus, $\mathcal{M}(I) = (O, R)$. Every row in this matrix $O \in \mathbb{R}^{NxN_o}$ presents an object proposal embedding n_j where $j \in 1, ..., N$. At the same time, each row in relation matrix $R \in \mathbb{R}^{(N*N-1)xN_r}$ refers to a relation distribution $r_{j\to i}^{t-1}$ between the node embeddings n_i and n_j. The object and edge embedding n_i, $r_{j\to i}^{t-1}$ are produced by applying the rows of O and R i.e. o_i & r_{ij} to a kernel MLP layer.

$$m_i^t = \alpha \underbrace{W_{nn}\left(\sum_{j\in N} n_j^{t-1}\right)}_{\text{node to node message}} + (1-\alpha) \underbrace{W_{en}\left(\sum_{j\in N'} r_{j\to i}^{t-1}\right)}_{\text{edge to node message}} \quad (7)$$

Equation 7 shows the message embedding vector for the i-th proposal. These messages are aggregated from the neighboring nodes and edges for each node. α denotes the weighting for the node and edge messages. The message passing helps the model to aggregate all relevant information treating every node related. Even if there is no information between the nodes, this information must be shared for the foreground objects.

$$\underbrace{p_{j\to i}^t = W_{ne}[n_i^{t-1}||n_j^{t-1}]}_{\text{node to edge message}} \quad (8)$$

A similar edge update message should be also computed from the neighboring nodes. However, the order of the nodes affects the resulting predicate so before sending the edge message, the nodes should be concatenated properly as it is shown in equation 8.

$$\begin{aligned} n_i^t &= GRU_n(m_i^t, n_i^{t-1}) \\ r_{j\to i}^t &= GRU_e(p_i^t, r_{j\to i}^{t-1}) \end{aligned} \quad (9)$$

To capture the composition in the image, the proposed energy model architecture leverages Gated Recurrent Units (GRUs). As in the modified Motif, GRUs cells aggregate the messages from neighboring nodes and edges. The hidden states of the GRUs are initialized from the node and edge embeddings. The reader may notice that there is a superscript t on the messages. The embeddings need to be updated in several iterations to get better node and edge states. That means the initial states are provided by the relation model.

Pooling Functions. Obtaining a scalar energy value requires a pooling operation on the updated node embeddings. Since one scalar value holds all the information present in the image, the authors in [13] suggests a clever way to aggregate all the nodes and edges. Equation 10 illustrates the attention-based pooling operation. Each node and edge is fed into a linear layer to get a particular attention score between 0-1. This attention score is multiplied with the embeddings to apply a weighted aggregation. The resulting vectors N, E represent the clear edge and node vector representations as it gives low attention to unrelated nodes and edges. Lastly, the concatenation of the N, E retrieves the scalar energy value for the input scene graph.

$$\begin{aligned} N &= \sum_k f_{gate}(n_k) \odot n_k \\ E &= \sum_{ij} g_{gate}(r_{i\to j}) \odot r_{i\to j} \\ e^- &= MLP(N; E) \end{aligned} \quad (10)$$

Sampling in Weak Supervision. As explained earlier, the message-passing algorithm is needed in the energy model, to refine the node and edge embeddings. However, this message passing should be only applied between the foreground objects to obtain valuable messages. Full supervision contains bounding box information so it knows foreground object locations. However, weak supervision does not allow you to know which proposals are the foreground proposals. This paper suggests that only the proposals that are in the image-level ground truth object labels should be considered in message-passing algorithms. Even though the relation model considers $N(N-1)$ object pairs, most of these pairs are unnecessary and create nothing but noisy messages. Therefore, removing some proposals in the energy model improves the model's performance. The detailed results are provided in the Section Ablation Studies IV-E.

The Handcrafted Background Score. Another problem that needs to be resolved is the background score. Section III-C provides the loss formulations for the combined model. The relation model is mainly trained with binary cross-entropy loss that uses image-level relation labels. However, this loss formulation does not train the model for producing a valuable background score since weak supervision does not know which pairs are unrelated. At the same time, the model still predicts meaningless scores for the background relation category. This paper recommends a handcrafted background score formulation as shown in equation 11.

$$\begin{aligned} BG &= 1 - \underset{r\in\mathbb{R}^{N_r}}{\text{argmax}}(softmax(R')) \in \mathbb{R}^{N(N-1)*1} \\ R &= [BG; R'] \end{aligned} \quad (11)$$

This formulation assigns a background score for the predicted relation score matrix R. The scores turned into probabilities using a softmax function. The background score BG for each object pair is then the subtraction of the maximum relation category probability for that pair from one. R' is the sliced version of the relation score matrix that does not contain meaningless background scores. Section Ablation Studies IV-E also considers the utilization of sigmoid instead of softmax.

C. Weakly Loss Formulations

To train the model a multi-loss function with three effective terms is formulated. Each loss function contributes to the overall performance of the model as follows: **i)** an image-level binary cross entropy loss $\mathcal{L}_{base}$ supervising the training regarding the image-level labels; **ii)** an energy loss that compares the predicted scene graph energy with the ground truth

image-level graph to incorporate the structure in the image; **iii)** a regularization loss for the calculated energy values. The formulation is the sum of all loss functions as shown in equation 12. The relative weights $\lambda_e, \lambda_{reg}, \lambda_{base}$ will be selected empirically.

$$\mathcal{L}_{total} = \lambda_e \mathcal{L}_e + \lambda_{reg} \mathcal{L}_{reg} + \lambda_{base} \mathcal{L}_{base} \tag{12}$$

The first term in the multi-loss function is named energy loss. The derivation of the energy loss is displayed in equation 13. The computation of the energy loss requires solving an optimization problem that minimizes the equation 13. This is often dealt with by solving this function iteratively using Stochastic Gradient Langevine Dynamics (SGLD) [23].

$$\mathcal{L}_e = E_\theta(G^+_{SG}) - \min_{G_{SG} \in SG} E_\theta(G_{SG}) \tag{13}$$

The intuition behind the energy loss is that the optimization updates the predicted scene graph nodes and edges while minimizing the energy of the prediction. These updates guide the training by assigning smaller energy values for the correct predictions. For the predictions that are not in the dataset, the energy values should be maximized to have a balance. This is also called contrastive divergence-based learning. Please see the Background Section of the thesis material to get more detailed information about these updates.

$$\mathcal{L}_{reg} = E^2_\theta(G^+_{SG}) + E^2_\theta(G_{SG}) \tag{14}$$

The writers of [13] argues that the scalar energy values become too large while training. Hence, they address this issue by adding an L2 regularization loss.

$$\mathcal{L}_{base} = -\sum_{c=1}^{C} \left(\mathbb{1}_{[c \in \mathcal{R}]} \log a + \mathbb{1}_{[c \notin \mathcal{R}]} \log 1 - a \right) \tag{15}$$

Finally, the base model loss is calculated by using a binary cross-entropy for the image-level relation labels. If a relation category is present in the ground truth image-level labels, the target probability is assigned as 1. If it is not present, the target should be equal to zero. The predicted image-level relation scores are computed by operating a maximization function on the relation score matrix R as shown in equation 16. The max operation on rows should produce predicted relation scores for the image.

$$\tilde{r} = \max_{rows} R \in \mathbb{R}^{N_r} \tag{16}$$

IV. Experimental Results

This section gives information about the experimental setup and results of the proposed method.

A. Experimental Setup

This section mentions the dataset split for the tests. The popular image dataset that contains relationship labels Visual Genome [10] is employed in the experiments. A pre-processed version of the Visual Genome by Xu et al. [17] is popular among the SGG community so it is utilized in all experiments. Visual Genome contains 108k images and the utilized split considers 150 object categories and 50 relation categories. In addition, this split leverages only 57,723 images in training, while 5000 images are kept separately for the validation set. The test set consists of 26,446 images meaning that the split on the whole set corresponds to a 70%-30% split for the training and test phases respectively.

B. Evaluation Metrics

This paper finds useful two conventional evaluation metrics for the weakly supervised SGG task. Recall@K (R@K) metric [16] is one of the oldest metrics for SGG. Although R@K is a biased metric due to the long-tailed distribution of the Visual Genome, it is still beneficial in weakly supervised SGG tasks. Another metric is the mean Recall@K (mR@K) introduced by [1]. mR@K eliminates the bias in R@K by taking the average of all predicates R@Ks. Hence, it shows the effect of the energy model if it improves the results for the rare predicates.

The evaluation metrics have to be calculated under a certain setting. **SGDet**: Scene Graph Detection is the scheme that is going to be shown in the results. SGDet expects to predict the scene graph in the given image I. The scene graph means that the bounding boxes of the object should be detected with IoU ≥ 0.5, and the labels of the objects must be correct. Moreover, the relations between these correct object pairs need to be identified correctly.

C. Implementation Details

Detector. A pre-trained Faster R-CNN [24] with ResNeXt-101-FPN backbone [25], [26] is utilized for obtaining the proposal regions, and the weights of these layers are frozen during the scene graph model training. The pre-training on Visual Genome is a violation for the weak settings but this paper aims to prove the concept of the effectiveness of the energy-based method in weak settings. The detector outputs 30 proposals with 151 object categories including the background object category.

Relation Model. The relation model is trained with a Stochastic Gradient Descent(SGD) optimizer that has a learning rate of 0.1. The learning rate increases linearly until 400 iterations so it has a warm-up stage. The learning reduces at every plateau region according to the validation results. The decay happens only if 3 consecutive validation did not improve the validation results (plateau region.) SGD also has a momentum of 0.9 and an effective batch size of 4 in every experiment. The model's weights are initialized randomly, and a weight decay of 0.0001 is applied. During the training phase, the incorporation of frequency bias information is utilized to improve performance. The context biLSTMs have a hidden dimension of 512. The main GPU in training was The NVIDIA RTX 3090.

Energy Model. The α in message passing is set as 0.5 to keep a balance between messages. The message update iterations only applied 3 times. In SGLD optimization steps, the node and edge states are scaled back to the range of [0, 1]

when each update is made for the node and edge states. The node and edge embeddings have a vector size of 512. The relative weights $\lambda_e, \lambda_{reg}, \lambda_{base}$ are selected as 1.

D. Quantitative Results

This subsection presents the quantitative results from the experiments.

Model	Method	Scene Graph Detection	
		R@20/50/100	mR@20/50/100
Motif [19]	CE	25.48/32.78/37.16	4.98/6.75/7.90
	EBM	- / - / -	- / - / -
Motif [13]	CE	25.62/32.97/37.41	5.07/6.91/8.12
	EBM	24.39/31.74/36.29	5.67/7.71/9.27
Motif(Ours)	CE	25.13/32.01/35.90	4.91/6.73/7.89
	EBM	25.13/32.05/35.86	5.22/7.12/8.41

TABLE I
Paper reproduction results for full supervision.

The first experiment contains full supervision results for the modified relation and modified energy models. Table I illustrates these modification results. See the Thesis writing for the detailed modification lists for the models. This test is mainly done to establish solid ground for the other experiments in the paper. In addition, it displays how the modified energy model improves the baseline relation model. The first row of Table I states the original Motif model results for full supervision. The second row shows the re-implementation results from [13]. Finally, the third row in Table I mentions our full supervision results. Our modified energy model still enhances the baseline results for the mR@20-50-100 metrics. For instance, in mR@100, our baseline predicts 7.89 points whereas the energy model improves this value by 0.52 despite removing one of the important branches in the original work of [13] namely the image graph. Our results also display that R@20-50-100 values did not seem to change for the modified energy model. When one compares the third row with the second row, one sees that reproduction provided similar improvements for the mR@20-50-100 values but R@20-50-100 values did not really been affected. The expectation was to get some decreases in the R@K values for the energy model as it diminishes the bias in training. To understand the behavior of the experiment, Figure 4a displays an in-depth analysis of the R@100 values for each predicate separately for full supervision. As it can be seen from Figure 4a, the energy model supported the prediction for the rare predicates such as 'eat,''carry,' and 'riding.' but it did not really decrease the most biased predicates like 'on' or 'has.' The averaged biased predicates like 'holding,''above,' and 'near.' seemed to be diminished from R@100 by a couple of points. The reproduction does not seem to be perfect as in the original work of [13] but it seems to be succeeding for the uncommon predicates. The reason behind not getting the perfect reproduction is that the effective batch size was selected 16 in the original work whereas this paper selects 4 due to computational availability and the removal of an important branch in the original work. Another important result to mention is that mR@K and R@K are in balance. If one uses a method to boost mR@K, the R@K metric should decrease automatically since it is a biased metric. Putting more emphasis on unbiased predicates should take the accuracy of biased predicates away. Since reproduction does not increase mR@K values as significantly as in the original work(0.52 points for mR@100.), R@K metric can stay stable. Same Figure 4a also shows that some predicates do not even have a bar which means they did not have any successful predictions. The removal of the image graph in the original work caused a decrease in the performance for these semantically meaningful predicates.

The second experiment reveals the main results of this paper since it shows the effectiveness of the energy model under weak supervision. It also consists of a row for the results of the state-of-the-art studies. Table II summarizes all these experiments. The first row reveals fully supervised training results of the original Motif; hence, it possesses the highest values in every metric. The second row displays the state-of-the-art in weakly supervised SGG. Lastly, the third row projects our modified Motif baseline and the energy model results. One should start from the third row to examine the results. The weakly supervised baseline model has comparable results with the first row. Even though only image-level labels are used in the training, the weakly supervised baseline presented concrete results. The addition of the energy model improves these baseline mR@20-50-100 values marginally. However, R@20-50-100 values seem to get a drop of 0.89, 0.84, and 0.39 respectively when activating the energy model. This drop in R@K is expected because the energy model should put more emphasis on non-frequent predicates. However, it was assumed to get more performance for the mR@K values in the energy model.

To understand why the effectiveness of the energy model is marginal, a detailed predicate analysis is employed in Figure 4b similarly. Figure 4b illustrates a trend having a decrease for the frequent predicates, and a boost for the uncommon predicates. This behavior is similar to the full supervision results in Figure 4a as it also improves the predicates on the tail, and slightly reduces the ones around the head of the distribution. For example, the semantic predicates such as 'standing on,' 'carrying,' 'eating,' and 'riding' have an increasing trend when the energy model is activated whereas the common ones like 'on,' 'has' have diminished in Figure 4b. Furthermore, the values for these trends are also quite similar to the ones in full supervision. For instance, the maximum improvement for the predicate 'eating' in full supervision corresponds to 0.08 points in R@100 when the energy model is activated. The enhancement in weak supervision is around 0.04 points in R@100 making the improvements very close in terms of values. Although Table II illustrates minor advancements with the energy model, the detailed analysis reveals that the energy models tackle down to the biased nature of SGG training

Model	Supervision	Scene Graph Detection					
		R@20	R@50	R@100	mR@20	mR@50	mR@100
Motif [13]	Full	25.62	32.97	37.41	5.07	6.91	8.12
VSPNet [4]	Weak	-	4.70	5.40	-	-	-
LSWS [5]	Weak	-	7.30	8.73	-	-	-
WSGM [8]	Weak	4.12	5.59	6.45	-	-	-
Motif(Ours)	Weak-CE	22.25	28.51	31.75	3.14	4.58	5.59
	Weak-EBM	21.36	27.67	31.36	**3.22**	**4.70**	**5.74**

TABLE II

Quantitative Results. The weak supervision test results for CE & EBM models, and comparison with other studies in literature.

which also proves the main goal of this paper.

The minor improvements in the energy model are assumed to be caused by several reasons:

i) Section III proposes a handcrafted background probability calculation but this approach seems to be sub-optimal for assigning a background score. This score along with softmax activation is proposed to prevent smooth inputs for the energy model because the energy model requires sharp object and edge distributions for its inputs. However, the inputs are still not optimal as they contain high scores for 2 relation categories sometimes, as it is observed from the trained model debug sessions afterward. That is why the energy model achieves only minor successes for mR@K results.

ii) Another debug observation for the energy model is that the usage of the off-the-shelf object detector causes node-to-node messages to stop early as the nodes are not updated during training. Typically, the energy model also helps the object detection process in full supervision; however, the object detector outputs constant node states to be used directly without refinement in weak supervision. In addition, some of these constant object detections are noisy and not eliminated in the training process. Since full supervision allows a refinement stage for the detected object, the noisy proposals are removed during training. The suggestion for future work here is to add a refinement stage for object detection by still supervising detection results weakly. This should allow for the energy model to help the node states too, and not to stop learning early for the node-to-node kernels.

Furthermore, the second row in Table II displays some weak supervision SGG results in the literature. Even though these studies utilize similar image-level weak supervision losses in training, they use off-the-shelves detectors pre-trained on different datasets than Visual Genome. Therefore, it would be unfair to compare the performance of this thesis with these studies. Another idea that explains the huge gap in Table II between the weakly supervised baseline model result and the state-of-the-art model is the utilization of frequency baseline mentioned in [16]. These baseline assigns calculated statistics from Visual Genome on predictions to put more bias towards some relations depending on the object pair labels. This is not used in weakly SGG studies.

E. Ablation Studies

In this section, the selected ablations have been applied to the proposed method to show their contribution to overall results. Motif (+Softmax) results in Table III shows the best energy model results discussed earlier, and Motif (Baseline) is the baseline results for the only cross-entropy model.

Model	Method	Scene Graph Detection	
		R@20/50/100	mR@20/50/100
Motif(Baseline)	CE	22.25/28.51/31.75	3.14/4.58/5.59
Motif (No sampling)	EBM	22.22/28.43/31.67	3.02/4.40/5.41
Motif (+fg/bg)	EBM	21.90/28.17/31.44	3.02/4.51/5.50
Motif (Sigmoid)	EBM	21.43/27.94/31.43	3.10/4.56/5.58
Motif (+Softmax)	EBM	21.36/27.67/31.36	**3.22/4.70/5.74**

TABLE III

The ablation studies. The summary of all contributions for the proposed methods.

Foreground & Background Sampling. This paper proposed a sampling technique for weak supervision to make the energy model functional. As one may recall that full supervision offers the foreground objects and relations. Thus, the message passing between these foreground objects is useful in the energy model. However, weak supervision does not allow you to know which proposals are foreground or background. If you apply your message passing between all the pairs, you get the result for Motif (No sampling) in Table III. As one can see the energy model seems harmful when you allow message passing between unnecessary nodes. The addition of weak sampling leads to the results for the Motif (+fg/bg) in Table III. In that case, the energy model results improved compared to the previous case but it still did not improve the Motif (Baseline).

Softmax or Sigmoid Inputs. Preventing smooth distributions for the inputs of the energy model is provided with this proposed method. Softmax inputs also introduce the handcrafted background score discussed previously. As you can see from Table III, Motif (+Softmax) achieves the best results for the mR@20-50-100 metrics. Having the best results for the mR@K also causes to get the lowest values for R@20-50-100 as expected since these two metrics are in balance. Another suggestion was to use the sigmoid function instead

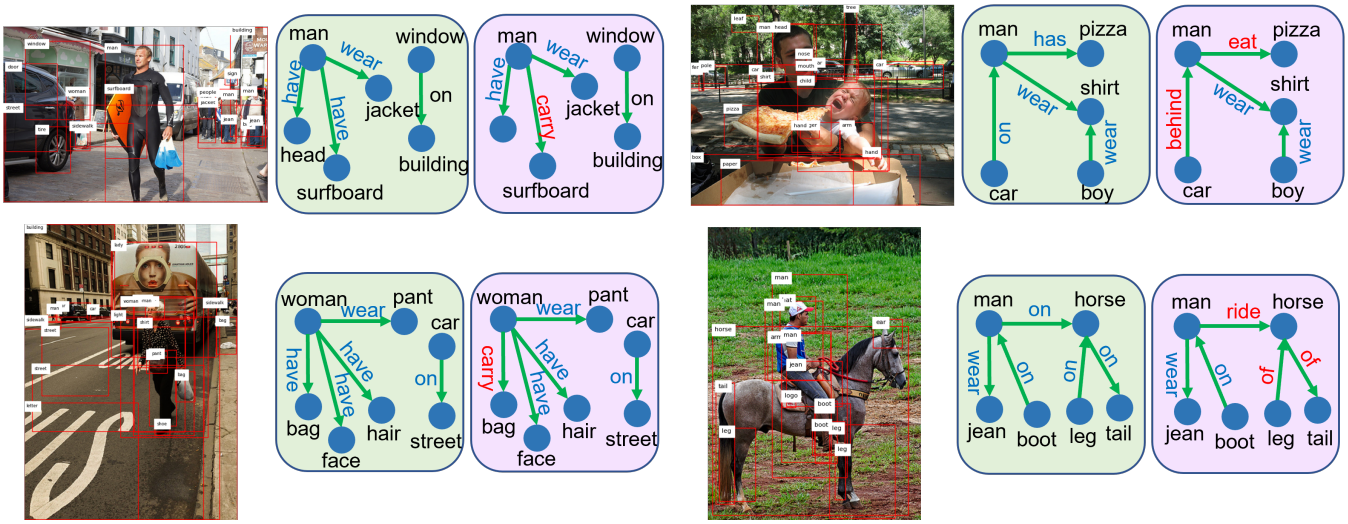


Fig. 3: **Qualitative Results.** Visualizations from the scene graph detection both for only cross-entropy (in green), and energy model (in purple)

of softmax. Even though it improved the results for the Motif (+fg/bg sampling), it is not as successful as softmax activation. The explanation of softmax is assumed to be caused by the sharp distribution requirements of the energy model. Sigmoid activation rewards every high relation score whereas softmax only rewards the highest relation category.

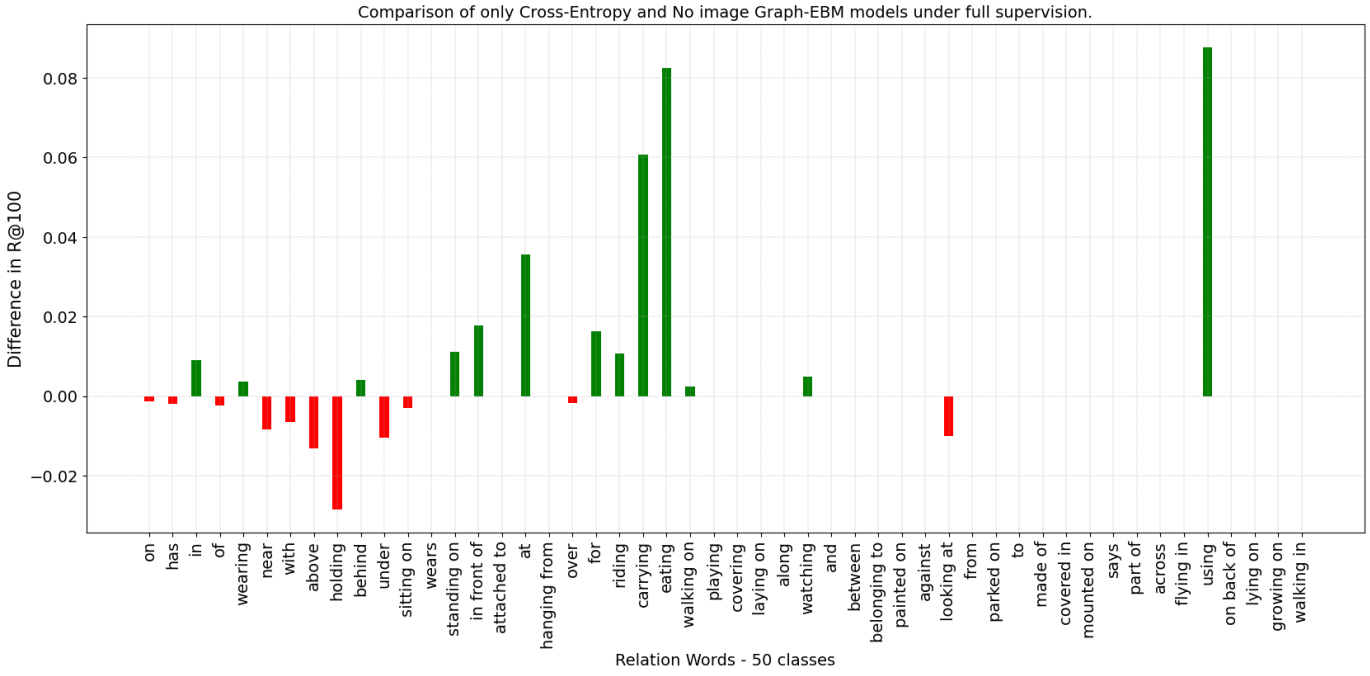
F. Qualitative Results

The performances of the models should be also compared to some example images. Figure 3 illustrates the best triplets for the given four example images. The results point out that the energy model is more successful in generating more instructive scene graph representations. One may recall the improvement trend in Section IV-A for the semantic predicates. These predicates could be observed easily in these example images. For instance, the energy model determined `<man eats pizza>` triplet instead of putting a more biased triplet such as `<man has pizza>`.

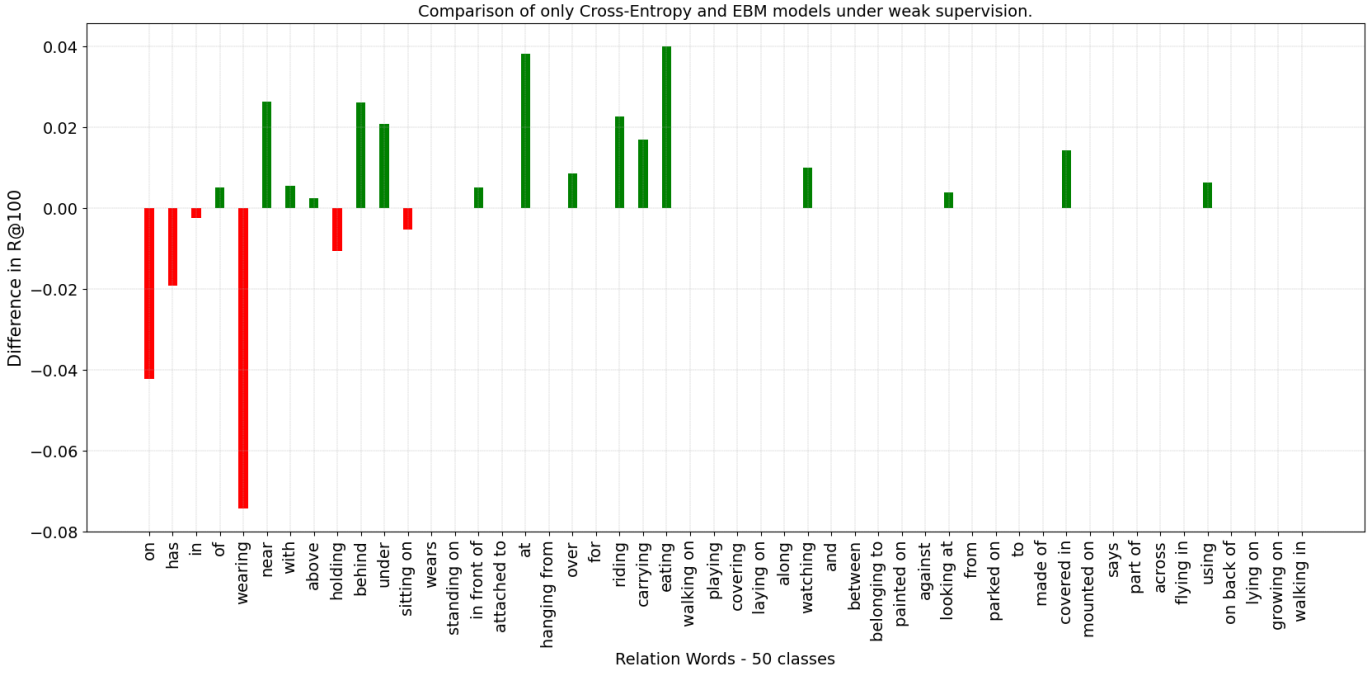
An intriguing finding from the qualitative results is that the images with obvious large objects lead to better performance for the energy model. For example, the relation `ride` is detected in the third image in Figure 3. This third image contains two large objects: 'man' and 'horse.' The energy model seems to achieve finding a better predicate when the detected objects are confident by the detector.

V. Conclusion

This paper suggested to benefits of energy-based approaches in weakly supervised SGG tasks. The experiment results showed that the modified relation model produces concrete results for SGG, and the energy approach improves the baseline model marginally. However, a detailed predicate analysis revealed that the improvements of the rare predicate in the dataset follow the same trend as in full supervision, proving the main goal of this paper. This work can be enhanced if one finds suggestions for the problematic parts of the method such as proposing a better background scoring method, and aggregation of additional information for the energy model instead of leveraging only from the object and relation distributions. In addition, the problem setup can be also extended to caption-based supervision to employ a weaker but more general strategy.



(a)



(b)

Fig. 4: In-depth analysis of each predicate word separately for the modified energy model under full supervision (a), weak supervision (b).

REFERENCES

[1] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019.

[2] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments, 2019.

[3] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[4] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8289–8299, 2021.

[6] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 612–630. Springer, 2020.

[7] Bo Wan, Yongfei Liu, Desen Zhou, Tinne Tuytelaars, and Xuming He. Weakly-supervised hoi detection via prior-guided bi-level representation learning. *arXiv preprint arXiv:2303.01313*, 2023.

[8] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16393–16402, 2021.

[9] Xingchen Li, Long Chen, Wenbo Ma, Yi Yang, and Jun Xiao. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4204–4213, 2022.

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[11] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[13] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13936–13945, 2021.

[14] Bo Pang and Ying Nian Wu. Latent space energy-based model of symbol-vector coupling for text generation and classification. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8359–8370. PMLR, 18–24 Jul 2021.

[15] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[16] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018.

[17] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[18] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[19] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020.

[20] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017.

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

[22] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.

[23] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

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

[25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

[26] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: [24/05/2023].