Iterative Scene Graph Generation

Siddhesh Khandelwal¹,² and Leonid Sigal¹,²,³
¹Department of Computer Science, University of British Columbia
²Vector Institute for AI
³CIFAR AI Chair
{skhandel, lsigal}@cs.ubc.ca

Abstract

The task of scene graph generation entails identifying object entities and their corresponding interaction predicates in a given image (or video). Due to the combinatorially large solution space, existing approaches to scene graph generation assume certain factorization of the joint distribution to make the estimation feasible (e.g., assuming that objects are conditionally independent of predicate predictions). However, this fixed factorization is not ideal under all scenarios (e.g., for images where an object entailed in interaction is small and not discernible on its own).

In this work, we propose a novel framework for scene graph generation that addresses this limitation, as well as introduces dynamic conditioning on the image, using message passing in a Markov Random Field. This is implemented as an iterative refinement procedure wherein each modification is conditioned on the graph generated in the previous iteration. This conditioning across refinement steps allows joint reasoning over entities and relations. This framework is realized via a novel and end-to-end trainable transformer-based architecture. In addition, the proposed framework can improve existing approach performance. Through extensive experiments on Visual Genome [30] and Action Genome [25] benchmark datasets, we show improved performance on the scene graph generation task. The code is available at github.com/ubc-vision/IterativeSG.

1 Introduction

Scene graphs allow for structured understanding of objects and their interactions within a scene. A scene graph is a graph where nodes represent objects within the scene, each detailed by the class label and spatial location, and the edges capture the relationships between object pairs. These relationships are usually represented by a <subject, predicate, object> triple. Effectively generating such graphs, from either images or videos, has emerged as a core problem in computer vision [14, 35, 39, 46, 50, 52, 54]. Scene graph representations can be leveraged to improve performance on a variety of complex high-level tasks like VQA [24, 47], Image Captioning [17, 53], and Image Generation [26].

The task of scene graph generation involves estimating the conditional distribution of the relationship triplets given an image. Naively modelling this distribution is often infeasible as the space of possible relationship triplets is considerably larger than the space of possible subjects, objects, and predicates. To circumvent this issue, existing methods factorize the aforementioned distribution into easy-to-estimate conditionals. For example, two-stage approaches, like [11, 47, 55], follow the graphical model image → [subject, object] → predicate, wherein the subjects and objects are independently obtained via a pre-trained detector like Faster-RCNN [41]. These are then consumed by a downstream network to estimate predicates. Any such factorization induces conditional dependencies (and independencies) that heavily influence model characteristics. For example, in the aforesaid graphical model, errors made during the estimation of the subjects and objects are naturally propagated towards the predicate distribution, which makes the estimation of predicates involving
classes with poor detectors challenging. Furthermore, the assumed fixed factorization might not always be optimal. Having information about predicates in an image can help narrow down the space of possible subjects/objects, e.g., predicate wearing makes it likely that the subject is a person.

Additionally, due to the two-stage nature of most existing scene graph approaches (barring very recent methods like [14, 31, 35]), the image feature representations obtained from a pre-trained task-oblivious detector might not be optimally catered towards the scene graph generation problem. Intuitively, one can imagine that the information required to accurately localize an object might not necessarily be sufficient for predicate prediction, and by extension, accurate scene graph generation. Moreover, two-stage approaches often suffer with efficiency issues as the detected objects are required to be paired up before predicate assignment. Doing so naively, by pairing all possible objects [30], results in quadratic number of pairs that need to be considered. Traditional approaches deal with this using heuristics, such as IoU-based threshold-based pairing [47, 55].

In this work, we aim to alleviate the previously mentioned issues arising from a fixed factorization and potentially limited object-centric feature representations by proposing a general framework wherein the subject, objects, and predicates can be inferred jointly (i.e., depend on one another), while simultaneously avoiding the complexity of exponential search space of relational triplets. This is achieved by performing message passing within a Markov Random Field (MRF) defined by components of a relational triplet (Figure 1(a)). Unrolling this message passing is equivalent to an iterative refinement procedure (Figure 1(b)), where each message passing stage takes in the estimate from the previous step. Our proposed framework models this iterative refinement strategy by first producing a scene graph estimate using a traditional factorization, and then systematically improving it over multiple refinement steps, wherein each refinement is conditioned on the graph generated in the previous iteration. This conditioning across refinement steps compensates for the conditional independence assumptions and lets our framework jointly reason over the subjects (s), objects (o), and predicates (p).

Contributions. To realize the aforementioned iterative framework, we propose a novel and intuitive transformer [48] based architecture. On a technical level, our model defines three separate multi-layer multi-output synchronized decoders, wherein each decoder layer is tasked with modeling either the subject, object, or predicate components of a relationship triplets. Therefore, the combined outputs from each layer of the three decoders generates a scene graph estimate. The inputs to each decoder layer are conditioned to enable joint reasoning across decoders and effective refinement of previous layer estimates. This conditioning is achieved implicitly via a novel joint loss, and explicitly via cross-decoder and layer-wise attention. Additionally, each decoder layer is also conditioned on the image features, which are provided by a shared encoder. As our proposed model is end-to-end trainable, it addresses the limitation of two-stage approaches, allowing image features to directly adopt to the scene graph generation task. Finally, to tackle the long-tail nature of the scene graph predicate classes [11], we employ a loss weighting strategy to enable flexible trade-off between dominant (head) and underrepresented (tail) predicate classes in the long-tail distribution. In contrast to data sampling strategies [11, 32], this has a benefit of not requiring additional fine-tuning of models with sampled data post training. We illustrate that our proposed architecture achieves state-of-the-art performance on two benchmark datasets – Visual Genome [30] and Action Genome [25]; and thoroughly analyze effectiveness of the approach as a function of the refinement steps, design choices employed and as a generic add-on to an existing, MOTIF [55], architecture.

2 Related Work

Scene Graph Generation. Scene graph generation has emerged as a popular research area in the vision community [11, 28, 31, 35, 39, 40, 46, 47, 50, 52, 54, 55]. Existing scene graph generation methods can be broadly categorized as either one-stage or two-stage approaches. The first step of the predominant approach – the two-stage methods – involves pre-training a strong object detector
for all object classes in the dataset, usually using detector like Faster-RCNN [41]. The graph generation network is then built on top of the object information (bounding boxes and corresponding features) obtained from the detector. This second step entails freezing the detector parameters and solely training the graph generation network. The graph generation network is realized via different architectures such as Bi-directional LSTMs [55], Tree-LSTMs [47], Graph Neural Networks [52], and other message passing algorithms [40, 50]. The limitation of all two-stage approaches is apparent – the graph generation network has no influence over the detector and object features. One stage approaches overcome this obstacle by employing architectures that allow end-to-end training, such as fully convolutional networks [35] and transformer based models [7, 14, 31, 43]. The joint optimization enables interaction between the detection and the graph generation modules, leading to better scene graphs. However, all the aforementioned methods operate under the assumption of a fixed factorized model, inducing conditional independencies that might not be ideal under all scenarios.

**Long-Tail Recognition.** The task of scene graph generation suffers from the challenge of long-tail recognition due to the combinatorial nature of relational triplets (both object and predicate distributions are skewed). Long-tail recognition, in general, is a well studied problem in literature. Data sampling is a popular strategy [3, 20, 21, 27, 37, 42, 58] wherein the training data is modified to either over-represent tail classes (oversampling) or under-represent the head classes (undersampling). Specific to the task of scene graph generation, [11, 32] have explored the use of data sampling strategies to improve performance on tail predicate classes. On the contrary, Loss re-weighting strategies [1, 9, 13, 27, 34] assign higher weights or impose larger decision boundaries for tail classes. [51] recently adopted this paradigm for the task of scene graph generation by proposing a re-weighting strategy based on correlations between predicate classes. Conceptually similar to [51], we propose a novel re-weighting strategy and illustrate its effectiveness both on its own, and in combination with data resampling.

**Transformer Models.** In [48], authors introduced a new attention-based architecture called transformers for the task of machine translation, doing away with recurrent or convolutional architectures. Transformers have since widely been adopted for a variety of tasks, such as object detection [2, 38, 57], image captioning [8, 23], and image generation [12]. More recently, scene graph generation methods have also adopted transformer architectures [7, 14, 31, 43], owing to their end-to-end trainable nature and parallelism. Perhaps conceptually closest, [14], which focuses on visual relation detection (VRD) and not scene graph prediction, leverages the composite nature of relationships to simultaneously decode an entire relationship triplet alongside its constituent subject, object, and predicate components. Unlike [14], we do not explicitly model subject-predicate-object "sum" composites, resulting in a simpler architecture, and leverage iterative refinement procedure that relies on a different factorized attention mechanism. In the latest, concurrent work, [31] generates a set of entities and predicates separately, and then utilizes a graph assembly procedure to match the predicates to a pair of entities. In contrast to our formulation, this precludes conditioning of entities on predicate estimates; therefore relying on accuracy of less contextualized predictions for subsequent pairing.

### 3 Formulation

For a given image \( I \), a scene graph \( G \) can be represented as a set of triplets \( G = \{ r_i \}_{i \leq n} = \{ (s_i, p_i, o_i) \}_{i \leq n} \), where \( r_i \) denotes the \( i \)-th triplet \( (s_i, p_i, o_i) \), and \( n \) denotes the total number of triplets. The subject \( s_i \) denotes a tuple \( (s_{i,c}, s_{i,b}) \), where \( s_{i,c} \in \mathbb{R}^n \) is the one-hot class label, and \( s_{i,b} \in \mathbb{R}^4 \) is the corresponding bounding box coordinates. \( n \) is the total number of possible entity classes in the dataset. The object \( o_i \) and predicate \( p_i \) can similarly be represented as tuples \( (o_{i,c}, o_{i,b}) \) and \( (p_{i,c}, p_{i,b}) \) respectively. Note that, \( p_{i,b} \) corresponds to the box formed by the centers’ of \( s_{i,b} \) and \( o_{i,b} \) as the diagonally opposite coordinates. Additionally, \( p_{i,c} \in \mathbb{R}^n \) represents the corresponding one-hot predicate label between the pair \( (s_i, o_i) \), where \( v \) is the total number of possible predicate classes in the dataset. Then the task of scene graph generation can be thought of as modelling the conditional distribution \( \text{Pr}(G | I) \). This distribution can be expressed as a product of conditionals,

\[
\text{Pr}(G | I) = \text{Pr}(\{s_i\} | I) \cdot \text{Pr}(\{o_i\} | \{s_i\}, I) \cdot \text{Pr}(\{p_i\} | \{s_i\}, \{o_i\}, I)
\]

(1)

where \( \{ \} \) denotes a set. For brevity, we omit explicitly mentioning the total number of set elements \( n \) throughout the paper. Existing approaches model this product of conditionals by making some underlying assumptions. For example, [11, 47, 55] assume conditional independence between \( s_i \) and \( o_i \) as they rely on heuristics to obtain the entity pairs. However, modelling the conditional in Equation...
For a given image, the model extracts features via a convolutional backbone and a transformer encoder. The individual components of a relationship triplet are generated using separate subject, object, and predicate multi-layer decoders. The inputs to each layer of the decoder is appropriately conditioned. For example, for the predicate decoder, the positional embeddings are conditioned on the outputs generated by the subject and object decoders (blue \textit{Attn} module) and the queries are conditioned on the previously generated graph estimate (orange \textit{Attn} module). The model is additionally implicitly conditioned and trained in an end-to-end fashion using a joint matching loss.

Contrary to existing works, our proposed formulation moves away from the “one-shot” generation ideology described in Equation 1. We instead argue for modelling the task of scene graph generation as an iterative refinement procedure, wherein the scene graph estimate at step $t$, $G_t$, is dependent on the previous estimates $\{G_{t'}\}_{t'\leq t}$. Formally, our aim is to model the conditional distribution $\Pr(G_t | \{G_{t'}\}_{t'\leq t}, I)$. Assuming Markov property holds, this can be conveniently factorized as,

$$
\Pr(G_t | G_{t-1}, I) = \Pr(s^t_i | G_{t-1}, I) \cdot \Pr(o^t_i | s^t_i, G_{t-1}, I) \cdot \Pr(p^t_i | o^t_i, s^t_i, G_{t-1}, I).
$$

(2)

Note that even though we assume the flow of information to be from $s^t_i \rightarrow o^t_i \rightarrow p^t_i$ for $G_t$, conditioning on the previous graph estimate $G_{t-1}$ allows the subject, object, and predicate predictors to jointly reason and update their beliefs, leading to better predictions. Additionally, the framework described in Equation 2 is model agnostic and can be implemented using any of the existing architectures.

## 4 Transformer Based Iterative Generation

As described in Section 3, our proposed iterative scene graph generation formulation is agnostic to the architectural choices used to realise the subject, object, and predicate predictors. In this section we provide a realization of the proposed formulation in Equation 2 using Transformer networks [48].

The choice of using transformer networks is motivated by their natural tendency to model iterative refinement behaviour under the Markov property, wherein each layer of the transformer decoder takes as input the output of the previous layer. Our novel end-to-end trainable transformer based iterative generation architecture builds on top of the DETR [2] framework, which is shown to be effective for the task of object detection. Our proposed model architecture is shown in Figure 2.

Given an image, our proposed approach first obtains corresponding image features using a combination of a convolutional backbone and a multi-layer transformer encoder, akin to DETR [2]. These image features are used as inputs to the subject, object, and predicate predictors, each implemented as
a multi-layer Transformer decoder. To generate a scene graph estimate \( \hat{G}^t \) at step \( t \) in accordance with Equation 2, the queries used in each predictor decoder are appropriately conditioned. For example, in the case of the predicate decoder, its input queries are conditioned on the subject \( \{s'_t\} \) and object \( \{o'_t\} \) estimates at step \( t \). Additionally, the input to each predictor decoder is infused with all decoder estimates \( \{s_{t-1}^i, p_{t-1}^i, \hat{o}_{t-1}^i\} \) from the previous step \( t - 1 \) via a structured attentional mechanism. The entire model is trained end-to-end, with a novel joint loss applied at each step \( t \) to ensure the generation of a valid scene graph at every level. This section describes these components in detail.

4.1 Image Encoder

Similar to DETR [2], for each image \( I \), our proposed architecture uses a deep convolutional network (like ResNet [22]) to obtain image level spatial map \( I \in \mathbb{R}^{c \times w \times h} \), where \( c \) is the number of channels, and \( w, h \) correspond to the spatial dimensions. A multi-layer encoder \( f_e \) then transforms \( I \) into a position-aware flattened image feature representation \( Z \in \mathbb{R}^{d \times w \times h} \), where \( d < c \).

4.2 Predictor Decoders

Our approach models each of the subject, object, and predicate predictors using a multi-layer transformer decoder [2], which are denoted by \( f_s \), \( f_o \), and \( f_p \) respectively. The \( t \)-th layer of each decoder, denoted as \( f_s^t \); \( x \in \{s, o, p\} \), is tasked with generating the step \( t \) scene graph \( G^t \). Therefore, at each step \( t \), the decoders output a set of \( n \) feature representations \( \{q_{s,t,i}^t\}; x \in \{s, o, p\} \), which are then obtained via a combination of self-attention between the input queries, and encoder-decoder attention across the encoder output \( Z \). These attention modules allow the decoder to jointly reason across all queries, while simultaneously incorporating context from the input image.

At a particular step \( t \), naively using the inputs \( \{q_{s,t-1}^t\} \) and \( \{p_{x,t-1}^t\} \) to generate \( G^t \) forgoes leveraging the compositional property of relations. As described Equation 2, for any arbitrary step \( t \), our proposed formulation entails two types of conditioning for better scene graph estimation. The first involves conditioning decoders on the step \( t \) outputs, specifically the object decoder \( f_o^t \) on the subject decoder \( f_s^t \), and the predicate decoder \( f_p^t \) on both \( f_s^t \) and \( f_o^t \). The second requires all three decoder layers at step \( t \) to be conditioned on the outputs generated at step \( t - 1 \). To effectively implement this design, we modify the inputs to each decoder layer \( f_x^t \). Specifically, the positional encodings \( \{p_{x,t-1}^t\} \) are modified to condition them on step \( t \) outputs, and the queries \( q_{x,t-1}^t \) are updated to incorporate information from the previous step \( t - 1 \). Modifying the positional encoding and queries separately allows the model to easily disentangle and differentiate between the two conditioning types.

Conditional Positional Encodings. At a particular step \( t \), the conditional positional encodings for the three decoders are obtained as,

\[
\begin{align*}
\{\hat{p}_{s,t,i}^t\} &= \{p_{s,t-1}^t\}, \quad \{\hat{p}_{o,t,i}^t\} = \{p_{o,t-1}^t\} + \text{FFN} \left( \text{MultiHead} \left( \{h_{o,t-1}^t\}, \{\hat{q}_{s,t,i}^t\}, \{\hat{q}_{o,t,i}^t\} \right) \right) \\
\{\hat{p}_{p,t,i}^t\} &= \{p_{p,t-1}^t\} + \text{FFN} \left( \text{MultiHead} \left( \{h_{p,t-1}^t\}, \{\hat{q}_{s,t,i}^t \oplus \hat{q}_{o,t,i}^t\}, \{\hat{q}_{s,t,i}^t \oplus \hat{q}_{o,t,i}^t \oplus \hat{q}_{p,t-1}^t\} \right) \right)
\end{align*}
\]

(3)

where \( \text{MultiHead}(Q, K, V) \) is the Multi-Head Attention module introduced in [48], \( \text{FFN}(.) \) is a fully-connected feed forward network, and \( \oplus \) is the concatenation operation. Additionally, \( \hat{q}_{x,t,i}^t = q_{x,t,i}^t + p_{x,t,i}^t; x \in \{s, o, p\} \) is the position-aware query.

Conditional Queries. Similarly, for a step \( t \), conditional queries for the subject decoder are defined,

\[
\{\hat{q}_{s,t-1}^t\} = \{q_{s,t-1}^t\} + \text{FFN} \left( \text{MultiHead} \left( \{\hat{q}_{s,t-1}^t \oplus \hat{q}_{o,t-1}^t \oplus \hat{q}_{p,t-1}^t\}, \{\hat{q}_{s,t-1}^t \oplus \hat{q}_{o,t-1}^t \oplus \hat{q}_{p,t-1}^t\} \right) \right)
\]

(4)

\( \hat{q}_{o,t-1}^t \) and \( \hat{p}_{o,t-1}^t \) are defined identically. For a decoder layer \( f_s^t; x \in \{s, o, p\} \) we use the conditioned positional encodings \( \{\hat{p}_{x,t,i}^t\} \) and queries \( \{\hat{q}_{x,t-1}^t\} \) as input.

4.3 End-To-End Learning

Our proposed transformer based refinement architecture can be trained in an end-to-end fashion. To ensure that a valid scene graph is generated at every level, we propose a novel joint loss that is applied at each step \( t \). Therefore, the combined loss \( L \) can be expressed as \( L = \sum_t L_t = \sum_t L_s + L_o + L_p \).
We demonstrate the effectiveness of our proposed approach on two datasets, Action Genome and Visual Genome, where (no relation). The goal then is to find a bipartite matching between the ground truth graph and the set of all graph estimates \( \hat{G}^t \) that minimizes the joint matching cost. Specifically, assuming \( \sigma \) to be a valid permutation of \( n \) elements,

\[
\hat{\sigma} = \arg \min_{\sigma} \sum \sum L_{\text{rel}} \left( r_i, \hat{r}_{\sigma(i)} \right)
\]

where the pair-wise relation matching cost \( L_{\text{rel}} \) is defined as,

\[
L_{\text{rel}} \left( r_i, \hat{r}_{\sigma(i)} \right) = -\mathbb{I}_{\{r_i \neq \emptyset\}} \left[ \hat{s}_{\sigma(i),c} \cdot \hat{s}_{i,c} - L_{\text{box}} \left( \hat{s}_{\sigma(i),b}, s_{i,b} \right) + \hat{o}_{\sigma(i),c} \cdot \hat{o}_{i,c} - L_{\text{box}} \left( \hat{o}_{\sigma(i),b}, o_{i,b} \right) + \hat{p}_{\sigma(i),c} \cdot \hat{p}_{i,c} - L_{\text{box}} \left( \hat{p}_{\sigma(i),b}, p_{i,b} \right) \right]
\]

where \( \cdot \) is the vector dot product, and \( L_{\text{box}} \) is a combination of the L-1 and generalized IoU losses. Please refer to Section 3 for clarification on the notations. The optimal permutation \( \hat{\sigma} \) can then be computed using the Hungarian algorithm. The loss \( L^t_s \) is then defined as,

\[
L^t_s = \sum_{i=1}^{n} \left[ -\log \left( \hat{s}_{\sigma(i),c} \cdot s_{i,c} \right) + \mathbb{I}_{\{r_i \neq \emptyset\}} L_{\text{box}} \left( \hat{s}_{\sigma(i),b}, s_{i,b} \right) \right]
\]

\( L^t_s \) and \( L^t_p \) are defined identically. Note that as we use the same permutation \( \hat{\sigma} \) for all refinement layers \( t \), it induces strong implicit dependencies between the subject, object, and predicate decoders. The potency of the aforementioned implicit conditioning is highlighted in the experiment section.

### 4.4 Loss Re-Weighting

Due to the inherent long-tail nature of the scene graph generation task, using an unbiased loss often leads to the model prioritizing the most common (a.k.a., head) predicate classes like has and on, which have abundant training examples. To afford our proposed model the flexibility to achieve the trade-off between head/tail classes, we integrate a loss-reweighing scheme into the model training procedure. Note that, contrary to existing methods that do this post-hoc via finetuning the final layer of the trained network (see Section 2), we instead train the model with this weighting to allow the internal feature representations to reflect the desired trade-off. Note, to the best of our knowledge, our paper is the first to illustrate effectiveness of such a strategy for the task of scene graph generation. For a particular predicate class \( c \in \{1, v\} \), we define the class weight \( w_c \) as \( \max \left\{ \alpha \cdot \beta^j, 1.0 \right\} \), where \( f_c \) is the frequency of the predicate class \( c \) in the training set, and \( \{\alpha, \beta\} \) are scaling parameters. Note that this weighting scheme is similar to the data sampling strategy described in [19, 32]. However, instead of modifying the training set, we scale each class weight by the factor \( w_c \) when computing the predicate classification loss \( L^t_p \). Therefore, \( L^t_p \) can be defined similarly to Equation 7.

\[
L^t_p = \sum_{i=1}^{n} \left[ -w_c \log \left( \hat{p}_{\sigma(i),c} \cdot p_{i,c} \right) + \mathbb{I}_{\{r_i \neq \emptyset\}} L_{\text{box}} \left( \hat{p}_{\sigma(i),b}, p_{i,b} \right) \right].
\]

### 5 Experiments

We demonstrate the effectiveness of our proposed approach on two datasets,

**Visual Genome** [30]. This is a benchmark for scene graph generation. We use the common processed subset from [50], which contains 108k images, with 150 object and 50 predicate categories.

**Action Genome** [25]. This dataset provides scene graph annotations over videos in the Charades dataset [44] for the task of human-object interaction. It contains 9,848 videos across 35 object
Table 1: Scene Graph Generation on Visual Genome. Mean Recall (mR@K), Recall (R@K), and Harmonic Recall (hR@K) shown for baselines and our approach. **B**=Backbone, **D**=Detector. Baseline results are borrowed from [31]. $M$ indicates the number of top-k links used by the baseline [31]. Note that our approach implicitly assumes $M = 1$. For the approaches that use the ResNet-101 backbone [22] (akin to our method), the best result is highlighted in red, second best in blue.

<table>
<thead>
<tr>
<th>B</th>
<th>D</th>
<th># Method</th>
<th>mR@50/100</th>
<th>R@50/100</th>
<th>hR@50/100</th>
<th>Head</th>
<th>Body</th>
<th>Tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ] RelDN [56]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>6.0 / 7.3</td>
<td>31.4 / 35.9</td>
<td>10.1 / 12.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] MOTIF [55]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>5.5 / 6.8</td>
<td>32.1 / 36.9</td>
<td>9.4 / 11.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] VCTree [47]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>6.6 / 7.7</td>
<td>31.8 / 36.1</td>
<td>10.9 / 12.7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] BGNN [32]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>10.7 / 12.6</td>
<td>31.0 / 35.8</td>
<td>15.9 / 18.6</td>
<td>34.0 / 12.9</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>[ ] VCTree-TDE [46]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>9.3 / 11.1</td>
<td>19.4 / 23.2</td>
<td>12.6 / 15.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] VCTree-DLFE [6]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>11.8 / 13.8</td>
<td>22.7 / 26.3</td>
<td>15.5 / 18.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] VCTree-EBM [45]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>9.7 / 11.6</td>
<td>20.5 / 24.7</td>
<td>13.2 / 15.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] VCTree-BPLSA [18]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>13.5 / 15.7</td>
<td>21.7 / 25.5</td>
<td>16.6 / 19.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] DT2-ACBS [11]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>22.0 / 24.4</td>
<td>15.0 / 16.3</td>
<td>17.8 / 19.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[ ] BGNN [32, 31]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>8.6 / 10.3</td>
<td>28.2 / 33.8</td>
<td>13.2 / 15.8</td>
<td>29.1</td>
<td>12.6</td>
<td>2.2</td>
</tr>
<tr>
<td>[ ] RelDN [56, 31]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>4.4 / 5.4</td>
<td>30.3 / 34.8</td>
<td>7.7 / 9.3</td>
<td>31.3</td>
<td>2.3</td>
<td>0.0</td>
</tr>
<tr>
<td>[ ] AS-Net [4]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>6.1 / 7.2</td>
<td>18.7 / 21.1</td>
<td>9.2 / 10.7</td>
<td>19.6</td>
<td>7.7</td>
<td>2.7</td>
</tr>
<tr>
<td>[ ] HOTR [29]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>9.4 / 12.0</td>
<td>23.5 / 27.7</td>
<td>13.4 / 16.7</td>
<td>26.1</td>
<td>16.2</td>
<td>3.4</td>
</tr>
</tbody>
</table>

(Concurrent Work)

| [ ] SGTR $M=1$ [31] | [ ] | [ ] | 12.0 / 14.6 | 25.1 / 26.6 | 16.2 / 18.8 | 27.1 | 17.2 | 6.9 |
| [ ] SGTR $M=3$ [31] | [ ] | [ ] | 12.0 / 15.2 | 24.6 / 28.4 | 16.1 / 19.8 | 28.2 | 18.6 | 7.1 |
| [ ] SGTR $M=3$,BGNN [32] [31] | [ ] | [ ] | 15.8 / 20.1 | 20.6 / 25.0 | 17.9 / 22.3 | 21.7 | 21.6 | 17.1 |
| [ ] Ours$_{(\alpha=0.0,\beta=\theta)}$ | [ ] | [ ] | 8.0 / 8.8 | 29.7 / 32.1 | 12.6 / 13.8 | 31.7 | 9.0 | 1.4 |
| [ ] Ours$_{(\alpha=0.14,\beta=0.5)}$ | [ ] | [ ] | 14.4 / 16.4 | 27.9 / 30.4 | 19.0 / 21.3 | 30.0 | 17.3 | 11.2 |
| [ ] Ours$_{(\alpha=0.07,\beta=0.75)}$ | [ ] | [ ] | 15.7 / 17.8 | 27.2 / 29.8 | 19.9 / 22.3 | 28.5 | 18.8 | 13.3 |
| [ ] Ours$_{(\alpha=0.14,\beta=0.75)}$ | [ ] | [ ] | 15.8 / 18.2 | 26.1 / 28.7 | 19.7 / 22.3 | 28.2 | 19.4 | 13.8 |
| [ ] Ours$_{(\alpha=0.14,\beta=0.75)}$,BGNN [32] | [ ] | [ ] | 17.1 / 19.2 | 22.9 / 25.7 | 19.6 / 22.0 | 24.4 | 20.2 | 16.4 |
| [ ] Ours$_{(\alpha=0.14,\beta=0.75),M=3}$ | [ ] | [ ] | 19.5 / 23.4 | 30.8 / 35.6 | 23.9 / 28.2 | 32.9 | 28.1 | 15.8 |

**Implementation Details (transformer-based approach).** We use ResNet-101 [22] as the backbone network for image feature extraction. Each of the subject, object, and predicate decoders have 6 layers, with a feature size of 256. The decoders use 300 queries. For training we use a batch size of 12 and initial learning rate of $10^{-4}$, which is gradually decayed. Note that although our model predicts individual relation triplets, a graph is obtained by applying a non-maximum suppression (NMS) strategy [15, 41] to group the predicted subjects and objects into entity instances.

**Implementation Details (MOTIF).** For the re-implementation of the MOTIF [55] baseline and the subsequent augmentation of our proposed framework, we follow the same training procedure as [55]. Specifically, we assume the Faster-RCNN detector [41] with the ResNetXt-101-FPN [49] backbone. The detector is first pre-trained on the Visual Genome dataset [30]. As is the case with two-stage approaches, when learning the scene graph generator, the detector parameters are freeze. The specifics on how our approach is augmented to MOTIF is described in the supplementary (Sec. A).

**Evaluation Metrics.** To measure performance, we report results using standard scene graph evaluation metrics, namely Recall (R@K) [50] and Mean Recall (mR@K) [5, 47]. While recall is class agnostic, mean recall averages the recalls computed for each predicate category independently. Usually, higher R@K is indicative of better performance on dominant (head) classes, where as higher mR@K suggests better tail class performance. Recent methods [11, 32] have argued for the use of mean recall as it reduces influence of dominant relationships such as on and has on the metric. However, even for the same model architecture, one can trade-off R@K for mR@K by using long tail recognition techniques described in Section 2. This trade-off is seldom analyzed or reported, making it difficult to compare performance across methods. Therefore, inspired by the generalized zero-shot
We vary all the subsequent ablation studies are done on the Visual Genome dataset. Table 4: Ablation of Model Components. We vary \( \alpha, \beta \) (Sec. 4.4) and report recall, mean recall, and harmonic recall on the Visual Genome test set.

<table>
<thead>
<tr>
<th>#</th>
<th>CAS</th>
<th>CWS</th>
<th>JL</th>
<th>mR@20/50</th>
<th>R@20/50</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \odot )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>11.8 / 15.8</td>
<td>21.0 / 26.1</td>
</tr>
<tr>
<td>( \Diamond )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>11.2 / 14.7</td>
<td>20.1 / 25.4</td>
</tr>
<tr>
<td>( \odot )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>11.5 / 15.3</td>
<td>20.9 / 26.1</td>
</tr>
<tr>
<td>( \odot )</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>11.7 / 15.4</td>
<td>20.9 / 25.9</td>
</tr>
</tbody>
</table>

Table 3: Augmentation to MOTIF. Mean Recall (mR@K) and Recall (R@K) is reported on the Visual Genome dataset for the baseline and the variant with our iterative formulation. ↑ denotes our re-implementation of the method.

<table>
<thead>
<tr>
<th>Model</th>
<th>t</th>
<th>mR@20/50</th>
<th>R@20/50</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOTIF↑</td>
<td>0.0 / 8.0</td>
<td>23.6 / 30.4</td>
<td></td>
</tr>
<tr>
<td>MOTIF↑ +</td>
<td>1.0 / 8.1</td>
<td>23.7 / 30.6</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>6.4 / 8.5</td>
<td>24.0 / 30.8</td>
<td></td>
</tr>
</tbody>
</table>

learning harmonic average metric [10], we propose a new metric – harmonic recall (hR@K), defined as the harmonic mean of mR@K and R@K. Such a metric encourages methods to have a healthy balance between the head and tail class performance. Additionally, we report the performance of our model for different values of the loss weighing parameters described in Section 4.4. Furthermore, we also report the mR@100 on each long-tail category subset, namely head, body, and tail, as in [32].

5.1 Ablation Study

All the subsequent ablation studies are done on the Visual Genome dataset [30].

Model Components. We analyze the importance of each of our model components, namely the conditioning within a particular step \( t \) (CWS; Equation 3), the conditioning across steps (CAS; Equation 4), and the joint loss (JL; Section 4.3) in Table 2. For a fair comparison, all models are trained using the same loss weighing parameters, \( \alpha = 0.14, \beta = 0.75 \). It can be seen that our proposed novel joint loss provides a significant improvement in performance (\( \odot \)) when compared to the ablated model that uses separate losses (as in [2]; \( \otimes \)). This is largely due to the joint loss being able to induce strong implicit dependencies between the three decoders. As a consequence, even without CAS and CWS, the proposed joint loss by itself enables refinement. The two forms of conditioning (\( \odot \)-\( \otimes \)) provide a structured pathway to incorporating knowledge from previously generated estimates or other triplet components in an explicit fashion, leading to improved scene graphs. CAS (\( \otimes \)) builds upon the joint loss, integrating information that the aforementioned implicit conditioning is unable to capture. CWS (\( \otimes \)) is complementary to the refinement process, and allows for more consistent graph generation within a step. As these two types of conditionings capture complementary information, using them together (\( \otimes \)) further improves on performance.

Refinement Across Steps. To further analyze the effectiveness of our refinement procedure, we contrast the quality of scene graphs generated at each refinement step for a particular trained model (\( \alpha = 0.14, \beta = 0.75 \)). We provide a detailed analysis in the supplementary (Table A1). Visually, Figure 3 highlights steady improvements in the graph quality over refinement steps. Owing to our structured conditioning, the model is able to improve on both predicate detection and entity localization over refinement iterations.
Loss Re-Weighting. As mentioned in Section 4.4, we employ a simple loss re-weighting scheme, allowing for an easy trade-off between recall for mean recall. Some results with different weightings are reported in Table 1 (F–G). To investigate this further, we train our proposed model with different values for the parameters $\alpha$ and $\beta$, and report the performance in Table 4. We fix $\beta$ in the top half of the table (first 4 rows), and $\alpha$ in the bottom half (last 4 rows). It can be seen that increasing $\alpha$ and $\beta$ generally biases the model more towards the tail classes (higher mean recall), whereas lower values tend to favour the head classes (higher recall). As a consequence, our proposed approach is flexibly able to function in a wide range of recall and mean recall values. However, as is the case with every long-tail recognition approach, choosing values that are too high (e.g., $\alpha = 0.21, \beta = 0.75$) lead to poor performance due to the model being extremely biased and therefore unable to learn accurate feature representations on the popular classes.

Model Parameters. We additionally argue that the improvements obtained by our proposed refinement procedure are not a direct consequence of having an increased number of parameters. To corroborate this, we contrast our proposed model with alternatives with similar number of parameters but devoid of refinement in Table 5. (\textcircled{1}) corresponds to our transformer-based approach with 6 decoder layers that uses the proposed joint loss and both forms of conditioning. (\textcircled{2}) is a similar model with 6 decoder layers but bereft of the aforementioned proposed implicit and explicit conditionings. (\textcircled{2}) on the other hand, is a single decoder layer transformer model without refinement, wherein the number of parameters are increased to be comparable to our proposed approach. The inferior performance of (\textcircled{2}) and (\textcircled{1}) highlight that increasing model capacity either via making each layer larger or via adding more layers (increasing model depth) does not emulate refinement. Our proposed refinement procedure, realized by the joint loss and explicit conditionings, allows for better learning, and the performance gains observed cannot be attributed to having additional parameters.

5.2 Comparison to Existing Methods

Visual Genome. We report the mean recall, recall, and harmonic recall values contrasting our proposed model with existing scene graph methods, including the concurrent work in [31], in Table 1. Contrary to existing approaches, depending on the loss weighting parameters used, our proposed approach is able to easily operate on a wide spectrum of recall and mean recall values (\textcircled{1}–\textcircled{2}). More concretely, compared to the most competitive baseline in SGTR [31] (\textcircled{6}–\textcircled{7}) and other one stage methods in AS-Net [4] (\textcircled{3}) and HOTR [29] (\textcircled{4}), our approach is able to achieve a considerably better recall on the head classes (\textcircled{5}: 3.5 mR@100 better on head classes compared to [31]). Additionally, when comparing a variant of our approach that has a similar R@100 value to SGTR (\textcircled{1} and \textcircled{2}), we do significantly better on mean recall (3.8 higher mR@50) highlighting the proficiency of our method to generalize to tail classes wherein the training data is limited. Furthermore, our proposed model in \textcircled{2} always achieves the best performance across all models (including two-stage methods that use a superior backbone [49]) on hR@50/100, underlining the capability of our method to perform well on both the head and tail classes simultaneously. SGTR [31] additionally reports numbers by utilizing a data sampling approach BGN [32] to bias the model towards the tail classes (\textcircled{5}). We highlight that, compared to (\textcircled{5}), our model in \textcircled{2} performs similarly on mR@50 but much better R@50/100 (5.5 R@50 higher). The poorer performance on mR@100 is largely due to SGTR [31] selecting top-3 subjects and objects for each predicate, leading to more relationship triplets being generated. On the contrary, our approach only uses 1 predicate per subject-object pair. For a fairer comparison, we evaluate our model in \textcircled{2} using a similar top-$k$ strategy, wherein we select the top-3 predicates for each subject-object pair. The resulting model (\textcircled{2}) outperforms the closest baseline by a significant margin (3.3 higher mR@100, 10.6 higher R@100 compared to \textcircled{5}). Additionally, we show that our loss re-weighting strategy is complementary to existing data sampling approaches fine-tuning our pre-trained model in \textcircled{2} using BGN [32] (\textcircled{2}). For further analysis, see supplementary Section C.2.
Table 6: Scene Graph Prediction on Action Genome. Mean Recall (mR@K), Recall (R@K), and Harmonic Recall (hR@K) shown for baselines and variants of our approach with different loss weighting parameters. B=Backbone, D=D detector. Baseline results are taken from [16].

<table>
<thead>
<tr>
<th>B/D</th>
<th>#</th>
<th>Method</th>
<th>R@20 / 50</th>
<th>mR@20/ 50</th>
<th>hR@20/50</th>
</tr>
</thead>
<tbody>
<tr>
<td>X101-FPN Faster R-CNN</td>
<td>1</td>
<td>VRD [36]</td>
<td>10.3 / 10.9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Freq Prior [55]</td>
<td>24.0 / 24.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>IMP [50]</td>
<td>23.9 / 25.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>MSDN [33]</td>
<td>24.0 / 25.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Graph R-CNN [52]</td>
<td>24.1 / 25.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>RelDN [56]</td>
<td>25.0 / 26.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>SimpleBase [16]</td>
<td>27.9 / 30.4</td>
<td>8.3 / 9.1</td>
<td>12.8 / 14.0</td>
</tr>
<tr>
<td>R-101 DETR</td>
<td>8</td>
<td>Ours$_{{a=0,\beta=*}}$</td>
<td>31.0 / 37.5</td>
<td>20.9 / 25.3</td>
<td>24.9 / 30.2</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>Ours$_{{a=0.07,\beta=0.75}}$</td>
<td>30.1 / 36.5</td>
<td>36.9 / 42.8</td>
<td>33.1 / 39.4</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Ours$_{{a=0.14,\beta=0.75}}$</td>
<td>29.2 / 35.3</td>
<td>37.9 / 44.0</td>
<td>32.9 / 39.2</td>
</tr>
</tbody>
</table>

**Action Genome.** We additionally report recall, mean recall, and harmonic recall values on the Action Genome dataset [25] in Table 6. Similar to the observations on Visual Genome, we observe 3.1 higher R@20 and 12.6 higher mR@20 ($\uparrow$) when compared to the closest baseline in [16] ($\downarrow$) despite using an inferior backbone [49]. Furthermore, by effectively biasing our model towards the tail classes, we are able to achieve 29.6 better mR@20 while having better R@20/50 values ($\uparrow$) compared to $\downarrow$.

We provide per class relation recall and object AP for both datasets in the supplementary (Section C).

5.3 Generality of Proposed Formulation

Although our transformer model effectively generates better scene graphs, the proposed iterative refinement formulation (Equation 2) can be applied to any existing method. To demonstrate this, we augment this framework to the simple two-stage MOTIF [55] architecture. The details on how this is achieved is deferred to the supplementary (Section A). We contrast the performance of the baseline and our augmented approach on the Visual Genome dataset [30] in Table 3. It can be seen that with each refinement step, the model is able to generate better scene graphs ($\sim$6% higher on mR@20 and mR@50 after 3 steps). Note that, owing to the two-stage nature of MOTIF [55], the image features used by our refinement framework are fixed and oblivious to the task of scene graph generation (as the detector parameters are frozen). Contrary to the proposed transformer model that can query the image at each layer, the performance gains when using the augmented MOTIF variant are largely limited due to the image features being the same across the refinement steps.

6 Conclusion, Limitations, and Societal Impact

In this work we propose a novel and general framework for iterative refinement that overcomes limitations arising from a fixed factorization assumption in existing methods. We demonstrate its effectiveness through a transformer based end-to-end trainable architecture.

**Limitations.** Limitations of the proposed approach include using a shared image encoder, which improves efficiency, but may limit diversity of representations available for different decoder layers; and limited ability to model small objects inherited from the DETR object detection design [31].

**Societal Impact.** While our model does not have any direct serious societal implications, its impact could be consequential if it is used in specific critical applications (e.g., autonomous driving). In such cases both the overall performance and biases in object / predicate predictions would require careful analysis and further calibration. Our loss re-weighting strategy is a step in that direction.

Acknowledgments and Disclosure of Funding

This work was funded, in part, by the Vector Institute for AI, Canada CIFAR AI Chair, NSERC CRC and an NSERC DG and Accelerator Grants. Hardware resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. Additional support was provided by JELF CFI grant and Compute Canada under the RAC award. Finally, we sincerely thank Bicheng Xu and Muchen Li for their valuable feedback on the paper draft.

1www.vectorinstitute.ai/#partners
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] see Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] see Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All the data we use in our experiments is publicly available. The code for our approach is available at github.com/ubc-vision/IterativeSG.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We use widely popular data splits for our experiments. We briefly describe this in Section 5. The hyperparameters and additional data details are also mentioned in the supplementary (Section B).
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] As scene graph models are computationally expensive to train, none of the existing literature report error bars. However, we do report the performance of our method on different hyperparameter values to enable better understanding of the performance improvements.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This is described in the supplementary (Section B).

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] All the datasets we use are publicly available. We cite the creators of Visual Genome (Krishna et al. [30]) and Action Genome (Ji et al. [25]) throughout the paper.
   (b) Did you mention the license of the assets? [Yes] Visual Genome [30] is licensed under the Creative Commons Attribution 4.0 International License. Action Genome [25] is licensed under the MIT license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We don’t introduce any new assets in the paper.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A] Both Visual Genome [30] and Action Genome [25] are publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]