Constrained Structure Learning for Scene Graph Generation

Daqi Liu, Miroslaw Bober, Member, IEEE, and Josef Kittler, Life Member, IEEE

Abstract—As a structured prediction task, scene graph generation aims to build a visually-grounded scene graph to explicitly model objects and their relationships in an input image. Currently, the mean field variational Bayesian framework is the de facto methodology used by the existing methods, in which the unconstrained inference step is often implemented by a message passing neural network. However, such formulation fails to explore other inference strategies, and largely ignores the more general constrained optimization models. In this paper, we present a constrained structure learning method, for which an explicit constrained variational inference objective is proposed. Instead of applying the ubiquitous message-passing strategy, a generic constrained optimization method - entropic mirror descent - is utilized to solve the constrained variational inference step. We validate the proposed generic model on various popular scene graph generation benchmarks and show that it outperforms the state-of-the-art methods.

Index Terms—Scene graph generation, structured prediction, mean field variational Bayesian, message passing, constrained optimization.

I. INTRODUCTION

The scene graph generation (SGG) task involves building a visually-grounded scene graph to explicitly model objects and their relationships in an input image. SGG provides a compact way to encode the input visual scene aiming to achieve a comprehensive scene understanding, which could also facilitate downstream vision tasks such as image captioning [1], [2] and visual question answering [3], [4]. It is essentially a visual context reasoning task, in which cognition is the core to such tasks that involve not just recognizing, but reasoning about our visual world. Current deep learning computer vision models achieve tremendous successes on simple visual perception applications, e.g., image classification or object detection. However, their performance still falls short to our expectations for such complex visual context reasoning tasks.

Manuscript received 16 August 2022; revised 7 April 2023; accepted 1 June 2023. Date of publication 5 June 2023; date of current version 5 September 2023. This work was supported in part by U.K. Defence Science and Technology Laboratory, in part by Engineering and Physical Research Council (collaboration between U.S. DOD, U.K. MOD, and U.K. EPSRC through the Multidisciplinary University Research Initiative under Grant EP/R018456/1, and in part by EPSRC under Grants MVSE (EP/V002856/1) and JADE2 (EP/T022205/1). Recommended for acceptance by X. Li. (Corresponding author: Daqi Liu.) The authors are with Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH Guildford, U.K. (e-mail: daqi.liu@surrey.ac.uk; m.bober@surrey.ac.uk; j.kittler@surrey.ac.uk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3282889, provided by the authors. Digital Object Identifier 10.1109/TPAMI.2023.3282889

As a structured prediction task, SGG is generally NP-hard, owing to the exponential complexity of interactions among the output variables (which are expected to form coherent visual relationships). It presents a huge challenge for directly computing the desired statistics, i.e., the underlying posterior or the relevant marginals. Currently, only pairwise interactions are considered in the SGG task and they are often formulated as triplet structures, in which each triplet consists of three components: a subject, a predicate and an object.

Visual relationship prediction can also be used to model the objects and their relationships. However, it can only produce independent and not consistent triplet structures. There are other rapidly emerging approaches to image interpretation, that are formulated as multi-modal and cross modal vision-language learning problems, such as image captioning [1], [2]. These could also potentially solve the SGG problem. These are out of the scope of the present paper, but investigating their relative merits vis-a-vis SGG is high on the agenda of future research. However, even if vision-language models become the methods of choice, the importance of modelling structural dependencies in the context of SGG is likely to inform the design of such multi-modal learning systems.

Given an input image \( x \), a specific type of approximation strategies - variational Bayesian (VB) [5], [6] - is often applied to accomplish the SGG generation task in the current methods. In this approach, the variational inference step aims to infer the optimum interpretation \( z^* \) by means of a max a posteriori (MAP) estimation strategy, i.e., \( z^* = \arg \max_{z} p(z|x) \), while the variational learning step tries to fit the model posterior \( p(z|x) \) with the underlying ground-truth posterior \( p_r(z|x) \) by maximizing the conditional likelihood. Such a VB framework is implemented in current SGG models [7], [8], [9], [10], [11], [12], [13], [14] by constructing two fundamental modules, namely: visual perception and visual context reasoning [15]. For the variational inference step, visual perception initializes the output interpretations, while visual context reasoning refines the interpretations according to certain inference strategies. For the variational learning step, both modules are updated to fit the ground-truth training samples, and the updated modules are applied in the following variational inference step. The resulting optimum interpretation \( z^* \) is generated by alternating between the above varitional inference and learning steps.

To construct efficient VB frameworks for complex SGG tasks, the variational distribution \( q(z) \) in the current SGG models is often assumed to be fully decomposable. The resulting framework is also known as mean field variational Bayesian (MFVB) [5].
Such formulation ignores the higher-order interactions in the underlying posterior and it is essentially a locally consistent (rather than globally consistent) approximation of the underlying posterior. However, the MFVB framework is easy to scale to huge datasets without sacrificing much performance, especially when used with stochastic learning methods. This explains why the majority of the current SGG methods choose this specific type of VB models as their backbone framework. In particular, almost all of the current MFVB-based SGG models adopt a specific optimization strategy - message passing [16], [17], [18], [19] - to infer the optimum interpretations within the variational inference step, and it has become the de facto inference method. Various message passing neural network structures [13], [14], [20], [21], [22] have been proposed in recent years to model the above MFVB models and showed to achieve reasonable graph generation performance.

Given an input image $x$, SGG aims to construct a visually-grounded output scene graph, in which the corresponding output variables $z$ are conditionally dependent. To solve the above structured prediction tasks, instead of directly enforcing the dependencies among the output variables $z$ like in classical conditional random fields (CRFs) [23], current SGG models rely on hidden conditional random fields (HCRFs) [24] by enforcing the dependencies among the latent continuous feature representations $y \in \mathbb{R}^d$ (where $d$ is the dimensionality). With such an implicit strategy, the current SGG models actually formulate an unconstrained optimization problem, since $y$ can take any values.

To accomplish the variational inference tasks in the above HCRFs, message passing techniques are often employed in the current SGG models. However, they may not be good candidates for the complex HCRFs used for modelling in SGG tasks, since they yield exact solutions only for a chain or a tree. More importantly, in the current SGG settings, the classical CRFs may be better than HCRFs to model SGG. This is because, in both, the end-to-end one stage SGG paradigm [25], [26], [27], [28], and the traditional two stages SGG paradigm [7], [14], [29], [30], the intrinsic structures of the output scene graphs (generally employed as the latent variables in traditional HCRFs) are often predefined.

In this paper, a generic and efficient constrained structure learning (CSL) method is proposed. Specifically, we formulate the SGG tasks as CRFs, in which the discrete output variables $z$ are conditionally dependent. The proposed CSL method is formulated using a constrained optimisation paradigm, in which the constraints are the variational distributions corresponding to the discrete output variables, confined to lie within a probability simplex. Such a paradigm makes it possible to apply advanced constrained optimization strategies (e.g., entropic mirror descent) to solve the inherent nonconvex optimization problems arising in SGG tasks. Thanks to the utilization of the geometry of the optimization problem, the entropic mirror descent strategy employed in the proposed CSL method improves the convergence and has a faster convergence rate [31], compared with the previous message passing techniques. The experimental results obtained on the popular Visual Genome and Open Images V6 benchmarks demonstrate the superiority and efficiency of the proposed CSL method.

The main contributions of the proposed CSL method can be summarized as follows:

1) The variational inference step in the proposed CSL method is formulated as a constrained optimization problem, rather than an unconstrained one in the previous SGG models.

2) In particular, the variational inference step aims to maximize an explicit evidence lower bound (ELBO) objective, subject to the constraint that the applied variational distribution resides in a probability simplex.

3) Instead of relying on the ubiquitous message passing strategy, the proposed CSL method employs a generic entropic mirror decent technique to solve the above formulated constrained optimization problem.

4) To our knowledge, the proposed CSL method is the first one to apply the above constrained optimization paradigm to efficiently solve the complex SGG tasks.

This paper is organized as follows: Section II presents related works. Section III introduces the proposed constrained structure learning methodology. The experimental results and the corresponding analysis are presented in Section IV. Finally, the conclusions are drawn in Section V.

II. RELATED WORKS

Current SGG models aim to find better feature extraction architectures [7], [8], [11], [20], [32], [33], or address the bias of the relationship prediction process, caused by the long-tail data distribution [13], [21], [29], [30], [34], [35]. With the exception of [30], which utilizes a causal inference, almost all of them tend to formulate the SGG task using a mean field variational Bayesian framework. In particular, the unconstrained variational inference objective is generally minimized by means of message-passing neural network structures, while the classical cross-entropy loss is often applied to train the associated learning frameworks. Such formulation has become a universal corner stone for almost all the current SGG tasks. In contrast, our proposed method presents an alternative SGG methodology, which constructs a constrained variational inference objective, and applies generic constrained optimization algorithms, rather than message-passing, to infer the optimum interpretation. It has been developed by investigating generic constrained optimization scenarios and by exploring the alternative inference strategies, which would further improve the applicability and diversity of the SGG methods.

Since the explicit variational inference objective is not required in message-passing based MFVB frameworks, the current SGG models do not need to specify the energy function or the scoring function for the input image $x$ and the output interpretation $z$. More specifically, given the input and output variables, energy function measures their dissimilarities, while scoring function gauges the corresponding similarities. Energy-based models (EBMs) [36], [37] aim to capture the dependencies among variables by associating a scalar energy to each potential configuration of the variables, which is generally non-probabilistic and can be converted to a probabilistic model, assuming the partition function can easily be computed or approximated. Such energy-based formulation is rarely
investigated in the current SGG literature and it is only explored by one recently proposed method [38]. However, the contrastive divergence loss applied in the above method may have mode collapse issue [39], which could underestimate the underlying posterior. Unlike the above energy-based algorithm, the proposed method approximates the associated partition function within the proposed MFVB framework.

From a broad perspective, scene graph generation is a type of structured prediction tasks, which naturally inherits its unique properties and solutions. Traditional techniques like Conditional Random Field (CRF) [23] or Structured Support Vector Machine (SSVM) [40] provide some basic ways to predict structured outputs $z^*$ from the input image $x$. However, these techniques are quite outdated in the current deep learning era. Therefore, several modern structured prediction methodologies [41], [42], [43] have been proposed in recent years, which leverage the powers of both classical structured prediction techniques and modern deep learning architectures. The representation learning capabilities of these techniques are greatly improved, which paves the way for extending them to more challenging applications. Following this direction, we propose a novel constrained structure learning methodology, which demonstrates its superior scalability and efficiency in complex SGG tasks.

III. PROPOSED METHODOLOGY

In this section we describe the proposed constrained structure learning method. It is organized as follows: Section III-A introduces the SGG problem formulation while Section III-B presents the applied scoring function. The variational Bayesian framework and the specific constrained variational inference strategy are discussed in the last two subsections. A graphical overview of the proposed method is presented in Fig. 1.

A. Problem Formulation

Given an input image $x$, a SGG model aims to build a visually-grounded scene graph by inferring the optimum coherent interpretation $z^*$ for all the objects and predicates within the input scene. Currently, only pairwise interactions are considered in the output scene graph, which consists of a list of intertwined semantic triplet structures, with each represented as $<s, p, o>$, where $s$ and $o$ are the associated subject and object, while $p$ is the corresponding predicate to describe the relationship between $s$ and $o$. In the current SGG approaches, the supporting evidence for the potential objects are captured by the associated bounding boxes, while their relationships are characterised by the observation in the corresponding union bounding boxes. The ground-truth training samples are represented as $(\hat{x}_i, \hat{b}_i, \hat{z}_i), i = 1, 2, \ldots, M$, where $M$ is the number of input images, $\hat{b}_i$ a list of ground-truth bounding boxes for potential objects in image $\hat{x}_i$, and $\hat{z}_i$ is a list of ground-truth labels for the objects and predicates in image $\hat{x}_i$.

To generate the underlying scene graph, two essential modules are required, namely, visual perception and visual context reasoning modules. The visual perception module aims to
locate and instantiate the potential objects and predicates within the input scene, while the visual context reasoning subsystem tries to infer the corresponding interpretations for these objects/predicates using certain inference strategies. In the current SGG tasks, a region proposal network (e.g., faster R-CNN [44]) with a VGG-16 [45] or ResNet-101 [46] backbone is often applied to implement the visual perception module, while the MAP inference is generally deployed to model the visual context reasoning module.

Given an input image \( x \), the aim of the visual perception module is to output a set of object region proposals \( b_i^o \in \mathbb{R}^4, i = 1, 2, \ldots, m \), as well as a set of predicate region proposals \( b_j^p \in \mathbb{R}^4, j = 1, 2, \ldots, n \), where \( m \) and \( n \) are the number of the potential objects and predicates within the input image, respectively. Specifically, suppose \( m \) objects are detected in an input image, a quadratic number of predicate proposals \( (m^2 - m) \) could potentially be generated by computing the pair of object proposal regions. In reality, the number of predicate proposals \( n \ll m^2 - m \) is much less, and the specific number is purely dependent on the underlying scene graph node adjacency structure of the ground-truth training samples. With the above region proposal sets, the input image \( x \) can be divided into two sets of image patches \( x_i^o, i = 1, 2, \ldots, m \) and \( x_j^p, j = 1, 2, \ldots, n \), respectively. Each of these image patches includes all the input pixels defined by its generating region proposals. A pooling strategy (e.g., ROI pooling) is applied to extract the corresponding feature representation sets \( y_i^o \in \mathbb{R}^d, i = 1, 2, \ldots, m \) and \( y_j^p \in \mathbb{R}^d, j = 1, 2, \ldots, n \).

Given a set of object classes \( C \) and a set of relationship types \( R \), a visual context reasoning module aims to infer a set of object labels (interpretations) \( z_i^o \in C, i = 1, 2, \ldots, m \) for the input image patch set \( x_i^o \), as well as a set of predicate labels \( z_j^p \in R, j = 1, 2, \ldots, n \) for the input predicate image patch set \( x_j^p, j = 1, 2, \ldots, n \).

With the above traditional SGG formulation, the global contextual information is largely ignored and only local contextual information is considered. In contrast, in our approach, a global latent feature representation \( y^\theta \) is incorporated into the proposed SGG framework. The associated input global image patch representation is denoted as \( z^\theta \). The corresponding global region proposal \( b^\theta \) is obtained by finding the unions of all the relevant objects and predicates within the input image. Given a set of global classes \( G \), although the corresponding interpretation set \( z^G \in G \) is not required, it is beneficial to incorporate such global contextual information, since it can consider higher-order interactions among the output variables.

### B. Scoring Function

Unlike the previous message-passing based SGG models, an explicit variational inference objective is required in the proposed method. To this end, a novel scoring function \( s_\theta(x, z) \) is used to model the similarity or compatibility between the input image \( x \) and the output scene graph \( z \). In particular, in this paper, neural networks (the feature representation learning functions \( h_\theta^o, h_\theta^p, g_\theta^{op}, g_\theta^{po}, g_\theta^{oo}, g_\theta^{po}, \rho_\theta^{op} \) in Appendix B, available online) are employed to construct the applied scoring function, in which \( \theta \) represents the weights and biases of the related neural networks. For an undirected graphical model, the scoring function can generally be represented as

\[
s_\theta(x, z) = \prod_{r \in R} f_r(x_r, z_r),
\]

where \( r \) is a clique within a clique list \( R \). \( f_r \) is a factor function describing the dependencies among the input image patch set \( x_r \) and the associated output interpretation \( z_r \).

In the current SGG tasks, only two types of factor functions are considered: the unary factor function \( f_u \) and the binary factor function \( f_b \). The former gauges the consistency between the input \( x \) and the label of a specific node, while the latter characterizes the interactions between a pair of nodes. In this paper, unless indicated otherwise, the discrete label \( z \) is generally represented as a corresponding one-hot vector, in which all the elements are set to zeros, except the one corresponding to the correct category.

To avoid a computationally intractable variational inference objective, \( f_r \) is generally formulated as an exponential function and the corresponding log scoring function becomes

\[
logs_\theta(x, z) = - \sum_{r \in R} \psi_r(x_r, z_r),
\]

where \( \psi_r \) is the corresponding potential function for the associated clique \( r \). Generally, SGG tasks have two types of potential functions: the unary potential function \( \psi_u \) and the binary potential function \( \psi_b \). The resulting posterior is computed as follows:

\[
p_\theta(z|x) = \frac{s_\theta(x, z)}{s_\theta(x)},
\]

where \( s_\theta(x) \) is the associated partition function or normalizing constant, and \( p_\theta(z|x) \) is essentially a Gibbs distribution. Specifically, in this paper, the log scoring function is defined in Appendix A, available online.

### C. Mean Field Variational Bayesian

The computational complexity of scene graph generation is generally NP-hard, since it is computationally intractable to integrate the exponentially growing number of structured outputs. For this reason, the existing SGG models tend to rely on a specific type of approximation strategies - variational Bayesian [5], [6] - to estimate the underlying posterior \( p_\theta(z|x) \) and infer the optimal interpretation \( z^* \) for an input image \( x \). A variational Bayesian (VB) model construction includes two alternating steps: variational inference and variational learning, in which the former aims to estimate the underlying posterior \( p_\theta(z|x) \) with a tractable variational distribution \( q(z) \), while the latter tries to fit the underlying posterior with the ground-truth data distribution \( p_r(z|x) \), i.e.,

\[
q^* = \arg \min_q \mathbb{D}(q(z), p_\theta(z|x))
\]

\[
\theta^* = \arg \min_\theta \mathbb{D}(p_r(z|x), p_\theta(z|x)),
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
where $\mathbb{D}$ is a divergence metric, normally chosen in the form of KL divergence. The optimum $q^*$ and $\theta^*$ are obtained by alternating between the above two divergence minimization steps, in which the first step performs variational inference, while the variational learning is executed in the second step.

However, it is impossible to infer the optimum $q^*$ by directly applying the first divergence metric minimization step, as it includes the computationally intractable posterior $p_\theta(z|x)$. Luckily, its dual problem - maximizing evidence lower bound (ELBO) - can be readily solved. Based on the Jensen’s inequality, the following equation can readily be derived:

$$
\log s_\theta(x) = \mathbb{E}_{q(z)} \log s_\theta(x, z) - \mathbb{E}_{q(z)} \log q(z) + \mathbb{E}_{p_\theta(z|x)} \log q(z) - \mathbb{E}_{q(z)} \log p_\theta(z|x),
$$

where, on the right-hand side, the first term is the so-called ELBO while the second term is the KL divergence between the variational distribution $q(z)$ and the underlying posterior $p_\theta(z|x)$. The term on the left-hand side is the log partition function, which is generally computationally intractable. Thus, maximizing ELBO has two consequences: 1) the KL divergence $\mathbb{D}(q(z), p_\theta(z|x))$ is minimized; 2) the resulting ELBO becomes a tighter lower bound of $\log s_\theta(x)$. Therefore, the maximization of ELBO is commonly applied to approximate the computationally intractable log partition function in the variational inference models.

For computational efficiency, the variational distribution $q(z)$ is generally assumed to be fully decomposed in the existing SGG models and each local variational distribution $q_i(z_i)$ is chosen from the conditionally conjugate exponential family [5] (categorical distribution for discrete output variables)

$$
q(z) = \prod_{i=1}^{m} q_i^*(z_i^*) \prod_{j=1}^{n} q_j^*(z_j^*),
$$

where $q_i^*(z_i^*) \in \Delta^{v_i - 1}$ and $q_j^*(z_j^*) \in \Delta^{v_p - 1}$ ($\Delta$ represents a probability simplex) are local variational distributions for the objects and predicates in the output scene graph, respectively. $v_o$ and $v_p$ are the sizes of vocabularies for the objects and predicates, respectively. With such an assumption, the resulting variational Bayesian model is also known as mean field variational Bayesian (MFVB) [5, 6], and the associated inference step is often called mean field variational inference (MFVI).

In MFVI, the indices of the maximal values of the marginals are exactly the same as the MAP inference results (which is not the case in general). Thus, the target MAP inference in SGG can be transformed into a corresponding marginal inference. In this paper, variable elimination technique [5] is applied to infer the associated marginals. Now, for a potential regional proposal $b_i$, delineating the input image patch $x_i$, the corresponding log marginal distribution is

$$
\log p_\theta(z_i|x_i) = \log s_\theta(x_i, z_i) - \log s_\theta(x_i) \\
\propto \sum_{\tilde{z}_i} \left[ \log s_\theta(x_i, z_i) \right] - \log s_\theta(x_i),
$$

where $\sum_{\tilde{z}_i}$ represents marginalization over the interpretations of all the potential output nodes, except the target node $i$, $\log s_\theta(x_i, z_i)$ is the associated log marginal scoring function, and $\log s_\theta(x_i)$ stands for the partition function associated with $x_i$. Specifically, the corresponding log marginal scoring function is defined in Appendix B, available online.

To infer the target log marginal $\log p_\theta(z_i|x_i)$, besides the above $\log p_\theta(x_i, z_i)$, it is necessary to estimate the computationally intractable $\log s_\theta(z_i)$. To this end, an explicit constrained variational inference objective is proposed

$$
\log s_\theta(x_i) \triangleq \max_{q_i} \mathbb{L}(q_i) = \max_{q_i} \mathbb{E}_{q_i(z_i)} \log s_\theta(x_i, z_i) q_i(z_i) \\
\quad \text{s.t. } q_i(z_i) \in \Delta^{v_i - 1},
$$

where $\Delta^{v_i - 1}$ is a $v - 1$ simplex and $\mathbb{L}(q_i)$ represents the variational inference objective. Unlike the previous SGG models, the variational inference step in the proposed method becomes explicit and is formulated as a constrained maximization problem. Specifically, the relevant constrained variational inference objective for a potential object/predicate region proposal can be found in Appendix C, available online.

Furthermore, the target log probability (or logit) $\log p_\theta(z_i|x_i)$ is computed via a surrogate logit $\phi$

$$
\phi = \log p_\theta(z_i|x_i) \triangleq \phi + C
$$

where $C$ is an associated constant w.r.t. $x_i$ and $z_i$. Using the LogSumExp trick, we can compute $\log p_\theta(z_i|x_i)$ by omitting the above constant $C$

$$
\log p_\theta(z_i|x_i) \triangleq \phi - \log \|e^\phi\|_1,
$$

where, for an input image patch $x_i$, its optimum interpretation is computed as $z_i^* = \arg\max_{z_i} \log p_\theta(z_i|x_i)$. For discrete output variables, $z_i^*$ is the index of the max value of the log probability $\log p_\theta(z_i|x_i)$.

To complete the proposed MFVB framework, a classical cross-entropy loss is applied to implement the associated variational learning step, which aims to fit $p_\theta(z|x)$ with the ground-truth training samples

$$
\theta^\star = \arg\min_\theta \mathbb{L}(\theta) = \arg\min_\theta \left[ \frac{1}{t} \sum_{k=1}^{t} \log p_\theta(z_k|x_k) \right],
$$

where $\mathbb{L}(\theta)$ represents the variational learning objective, $t$ is the number of training images in a mini-batch, $z_k$ is the ground-truth scene graph of the input image $x_k$.

D. Entropic Mirror Descent Inference Method

Unlike the previous message-passing based SGG models, the variational inference step in the above proposed MFVB framework is a constrained optimization problem, as shown in (8). Specifically, the variational inference step aims to maximize the associated ELBO $\mathbb{L}(q_i)$, subject to the constraint that the applied variational distribution $q_i(z_i)$ resides in a $v - 1$ probability simplex.
Algorithm 1: Entropic Mirror Descent Inference Method.
\textbf{Input} variational distribution \(q_i\), number of iterations \(T\), an initial learning rate \(\alpha\), a predefined objective \(\mathbb{L}_p(q_i)\), a small positive value \(\epsilon\).
\textbf{Output} optimum \(q^*_i\):
1: randomly initialize \(q_i\).
2: for iteration \(i = 1 \rightarrow T\) do
3: compute \(L(q_i)\) and its derivative \(\nabla_q L(q_i)\).
4: set learning rate \(\alpha = \alpha / \sqrt{t}\).
5: end the loop if \(|L(q_i) - \mathbb{L}_p(q_i)| < \epsilon\).
6: compute \(r = \alpha \cdot \nabla_q \mathbb{L}(q_i)\).
7: compute \(r = q_i \cdot e^{-\max(r)}\).
8: set \(q_i = \frac{r}{\|r\|_1}\).
9: end for.

The projected gradient descent (PGD) methods [47] are often applied to solve the above constrained optimization problem. Compared with the traditional gradient descent method, it essentially adds a \(L_2\) regularization term in the weight update step, which projects the updated weight to a valid set defined by the constraints. Mirror descent (MD) [48], [49] is a generalized projected gradient descent method in the sense that it replaces the above \(L_2\) euclidean distance with a more general Bregman distance [50]. Since the constraint in the above maximization problem is a probability simplex, the negative entropy \(\mathbb{E}_{q(z_i) \log q(z_i)}\) can be used as a specific function to construct the associated Bregman distance. The resulting algorithm is also known as the entropic mirror descent (EMD) [51].

The above generic entropic mirror descent method is applied to solve the associated constrained optimization problem formulated in the proposed variational inference step. The proposed entropic mirror descent inference method is summarised in Algorithm 1. Compared with the projected gradient descent algorithms, such method generally converges faster due to the utilization of the geometry of the optimization problem [31], which is especially desirable in complex SGG tasks.

IV. EXPERIMENTS
To validate the proposed method, in this section, it is compared with various state-of-the-art models on two popular scene graph generation benchmarks: Visual Genome [52] and Open Images V6 [53], respectively. An experimental analysis and ablation study are also presented. Finally, visualization results are provided and discussed in the last subsection.

A. Visual Genome

1) Benchmark: Visual Genome [52] is a predominant SGG benchmark, which contains 108,077 images with an average of 38 objects and 22 relationships per image. We adopt the same data split protocol as [7], in which the most frequent 150 object classes and 50 predicate classes are chosen for the experiment. Specifically, Visual Genome is divided into a training set (70%) and a test set (30%). An evaluation set (5 \(k\)), used for validation, is a random subset of the training set. Moreover, as in [54], based on the number of instances in training split, the categories are divided into three disjoint sets: \(head\) (more than 10 \(k\), bodies (0.5 \(k\) \(\sim\) 10 \(k\)) and \(tail\) (less than 0.5 \(k\)).

2) Evaluation Metrics: In this paper, as the evaluation metric we choose the mean Recall@\(K\) \((mR@K)\) rather than the regular Recall@\(K\) \((R@K)\), due to the data imbalance that leads to a bias, as demonstrated in [30]. In particular, \(R@K\) only focuses on common predicates (e.g., on), with abundant training samples, and underestimates the informative predicate categories (e.g., standing on or parked on) represented by a fewer training samples. Like the previous algorithms, we validate the proposed method on the following three settings: 1) Predicate Classification (PredCls), which predicts the predicate labels, given the input image, the ground-truth bounding boxes and object labels; 2) Scene Graph Classification (SGCls), which predicts the labels for objects and predicates, given the input image and the ground-truth bounding boxes; 3) Scene Graph Detection (SGDet), which predicts the scene graph from the input image. Lastly, we evaluate the relevant SGG models with, or without graph constraint: whether only a single relation with the highest confidence is predicted for each object pair. In particular, \(mR@K\) represents the evaluation metric with graph constraint, while \(mR@K\) depicts the one without graph constraint.

3) Implementation Details: As in [30], in this paper, ResNeXt-101-FPN [46] and Faster-RCNN [44] are chosen as the backbone and the object detector, respectively for the visual perception module. We choose the step training strategy, in which the pre-trained optimum parameters are loaded into the above models and kept frozen during training. To achieve an effective trade-off between the head and the tail categories, we adopt the same bi-level data resampling strategy as in [14], which includes image-level over-sampling (the data sampler creates a random permutation of images in which each image is repeated according to its repeat factor \(t\) in each epoch) and instance-level under-sampling (the data sampler under-samples based on a drop-out probability for instances of different predicate classes in each image). We set the repeat factor \(t = 0.07\) and the instance drop rate \(\gamma_d = 0.7\) in this experiment. The batch size \(bs\) is set to 12. For the PredCls and SGCls settings, we apply a two-layer MLP to construct the associated log scoring function and use a higher learning rate (0.008 \(\times\) \(bs\)) in the SGD optimizer. For the SGDet setting, we employ a three-layer MLP to build the corresponding log scoring function, and utilize a lower learning rate (0.005 \(\times\) \(bs\)) in the SGD optimizer.

4) Comparisons With State-of-the-Art Methods: Within the two blocks of Table I, we compare state-of-the-art SGG models using the end-to-end one stage paradigm and the traditional two stages paradigm, respectively. Specifically, in current SGG settings, the structures of the output scene graphs are often predefined and they can be considered as the prior knowledge for the corresponding variational inference framework. This applies to both end-to-end one stage paradigm or the traditional two stage paradigm. Specifically, to obtain the structure (could be fully connected or sparse) of the output scene graph, the end-to-end one stage paradigm uses the so-called queries while the traditional two stage paradigm employs an additional object detection stage.
As structured prediction tasks, SGG applications are required to accomplish the relevant variational inference steps to generate the consistent output scene graphs for the input images. In other words, as long as we consider the dependencies within the output scene graph, a variational inference step is required for the relevant SGG model. More importantly, besides the early one stage SGG models, which do not consider the dependencies among the output variables, for all the current SGG models (whether they are the end-to-end one stage SGG models or the traditional two stage SGG models), message passing is the only available solution to accomplish the variational inference tasks. In particular, for the end-to-end one stage SGG models, the fusion procedure employed to refine the relation features by using the relevant object features is essentially a message passing process.

As shown in Table I, the proposed CSL method achieves state-of-the-art performance in the SGCls and SGDet settings, and comparable performance with the latest BGNN model [14] in the PredCls setting. Specifically, compared with the latest BGNN method, the SGDet performance gain achieved by the proposed method is 11.2% and 13.5%, respectively. In our experiments, the proposed CSL method achieves the speed of 0.13 s per image. It is worth noting that the proposed CSL method can achieve such performance with a relatively small number of training iterations, since the generic EMD method applied in MFVI converges faster than the message passing strategy. For instance, in the SGDet task, the proposed CSL method converges at around 10,000 iterations while the BGNN model requires around 20,000 iterations to converge. Moreover, we also compare the proposed CSL method with various state-of-the-art SGG models using the evaluation metric $ng - mR@K$, as demonstrated in Table II. The proposed CSL method outperforms the previous SGG models by a large margin, especially in the PredCls setting. This further verifies the superiority of the proposed CSL method.

Furthermore, in Table III, we compare the performances on the long-tail category groups in the SGDet setting. Note, the proposed CSL method achieves the best mean performance. More importantly, CSL outperforms the previous methods by a large margin on the tail group, which clearly demonstrates its superior detection capability for the informative predicate categories with a fewer training samples. In other words, unlike the previous models, which mainly detect the dominant predicate categories, the proposed CSL method has the capacity to detect more informative predicate categories and thus reduce the problem of bias in the relationship prediction caused by the long-tail data distribution, as demonstrated in Fig. 2.

To improve the performance further, we adopt the generic balance adjustment strategy [35] into our proposed CSL method and compare the resulting performance with several state-of-the-art models in Table IV. For a fair comparison, we choose the three baseline models presented in [35]. The balance adjustment strategy includes two important processes: semantic adjustment and balanced predicate learning. The former aims to cast the common predictions generated by an SGG model as informative ones, while the latter tries to extend the sampling space for the informative predicates. These processes are applied to solve two

---

### Table I

| Method | PredCls | SGCls | SGDet |
|--------|---------|-------|-------|
|        | $mR@50$ | $mR@100$ | $mR@50$ | $mR@100$ | $mR@50$ | $mR@100$ |
| FCGG [35] | 6.3 | - | 3.6 | - |
| RoIR [26] | 21.2 | - | 10.8 | - |
| SSR-CNN [27] | - | - | - | - | 8.6 | 10.3 |
| RILDN [33] | 15.8 | 17.2 | 9.3 | 9.6 | 6.0 | 7.3 |
| Motifs [29] | 14.6 | 15.8 | 8.0 | 8.5 | 5.5 | 6.8 |
| MoDe [29] | 18.5 | 20.0 | 11.1 | 11.8 | 8.2 | 9.7 |
| G-RCNN [11] | 16.4 | 17.2 | 9.0 | 9.5 | 5.8 | 6.6 |
| MSDD [8] | 15.9 | 17.5 | 9.3 | 9.7 | 6.1 | 7.2 |
| VCTree [13] | 15.4 | 16.6 | 7.4 | 7.9 | 6.6 | 7.7 |
| GPS-Net [22] | 15.2 | 16.6 | 8.5 | 9.1 | 6.7 | 8.6 |
| GPS-Net+ [22] | 19.2 | 21.4 | 11.7 | 12.5 | 7.4 | 9.5 |
| Transformer [53] | 16.3 | 17.6 | 10.1 | 10.7 | 8.1 | 9.6 |
| VCTree-TDE [30] | 25.4 | 28.7 | 12.2 | 14.0 | 9.3 | 11.1 |
| BGNN [14] | 30.4 | 32.9 | 14.3 | 16.5 | 10.7 | 12.6 |
| CSL | 29.5 | 31.6 | 16.7 | 17.9 | 11.9 | 14.3 |

*Note: All the above methods apply ResNetXt-101-FPN as the backbone.

### Table II

| Method | PredCls | SGCls | SGDet |
|--------|---------|-------|-------|
|        | $ng@50$ | $ng@100$ | $ng@50$ | $ng@100$ | $ng@50$ | $ng@100$ |
| Motifs [29] | 32.8 | 42.7 | 33.5 | 42.7 |
| Motifs+ [27] | 29.8 | 35.2 | 31.2 | 37.2 |
| MoDe-TDE [30] | 29.0 | 38.2 | 31.2 | 37.2 |
| MoDe-CPL [38] | 4.8 | 25.6 | 14.8 | 19.6 |
| VCTree [13] | 35.6 | 47.8 | 35.6 | 47.8 |
| VCTree-TDE [30] | 31.8 | 40.1 | 32.0 | 40.2 |
| VCTree-CPL [36] | 34.0 | 41.3 | 35.4 | 42.7 |
| CSL | 36.0 | 51.9 | 26.6 | 32.9 |

*Note: All the above methods apply ResNetXt-101-FPN as the backbone.

---

### Table III

| Method | Head | Body | Tail | Mean |
|--------|------|------|------|------|
| RILDN+ [33] | 34.1 | 6.6 | 1.1 | 13.9 |
| Motifs [29] | 36.1 | 7.0 | 0.0 | 14.4 |
| MoDe+ [29] | 34.2 | 8.6 | 2.1 | 15.0 |
| G-RCNN+ [11] | 28.6 | 6.5 | 0.1 | 11.7 |
| MSDD+ [8] | 35.1 | 5.5 | 0.0 | 13.5 |
| VCTree-TDE [30] | 24.5 | 13.9 | 0.1 | 12.8 |
| GPS-Net+ [22] | 34.5 | 7.0 | 1.0 | 14.2 |
| GPS-Net+* [22] | 30.4 | 8.5 | 3.8 | 14.2 |
| BGNN [14] | 33.4 | 13.4 | 6.4 | 17.7 |
| CSL | 33.6 | 13.5 | 8.8 | 18.6 |

*Note: All the above methods apply ResNetXt-101-FPN as the backbone.

---

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
constrained structure learning for scene graph generation

Fig. 2. The mean Recall@100 performance (denoted by black dot) for each predicate category achieved by our proposed CSL method. The y-axis represents the min-max normalized frequency. For the tail categories (blue bars) with a fewer training samples, the proposed CSL method still achieves a reasonable detection rate, which demonstrates its ability to rectify the biased relationship prediction problem caused by long-tail data distribution.

TABLE IV
THE PERFORMANCE COMPARISON ON THE VISUAL GENOME DATASET USING THE BALANCE ADJUSTMENT STRATEGY

| Method      | PredCls mR@50 | PredCls mR@100 | SGCls mR@50 | SGCls mR@100 | SGGDet mR@50 | SGGDet mR@100 |
|-------------|---------------|----------------|-------------|--------------|--------------|---------------|
| Motifs+BA [38] | 29.7          | 31.7           | 16.5        | 17.5         | 13.5         | 15.6          |
| VCTree+BA [35] | 30.6          | 32.6           | 20.1        | 21.2         | 13.5         | 15.7          |
| Transformer+BA [36] | 31.9          | 34.2           | 18.5        | 19.4         | 14.8         | 17.1          |
| CSL+BA      | 36.9          | 39.2           | 19.7        | 21.2         | 15.7         | 18.4          |

* Note: All the above methods apply the same balance adjustment strategy as in [35].

Fig. 3. The mean Recall@100 performance (denoted by black dots) for each predicate category obtained with our proposed CSL+BA method, in which the y-axis represents the min-max normalized frequency. Based on the Shannon information theory, the informative tail (blue bars) and body (green bars) predicate categories are largely kept, while the number of samples of the common head (red bars) predicate categories are strictly controlled. Compared with Fig. 2, the training samples within the CSL+BA method are more balanced.

sub-problems: semantic space imbalance and training sample imbalance.

As shown in Table IV, the resulting CSL+BA method achieves the state-of-the-art performance on the Visual Genome benchmark. It outperforms the previous models by a large margin, especially for the PredCls setting. As demonstrated in Fig. 3, due to the balanced predicate learning, the resulting CSL+BA method has more balanced training samples, in which the more informative (from Shannon information theory perspective) tail and body predicate categories are largely kept, while the common head predicate categories are strictly constrained by means of training sample pruning. With the transition matrix introduced in the semantic adjustment process, the resulting CSL+BA method tends to choose the informative predicates rather than the common ones. In Fig. 4, we compare the R@100 performance of three SGG models (the baseline BGNN, the proposed CSL and the derived CSL+BA) on SGGDet task of all predicate categories. Compared with the baseline BGNN model, the proposed CSL method and the derived CSL+BA algorithm consistently achieve better detection performances on the informative body and tail predicate categories.

5) Ablation Study: In this section, we investigate the detection performance dependency of the proposed CSL method on the number of iterations $T$ of the entropic mirror descent optimisation procedure, and present the results in Table V. Note, the associated positive value $\epsilon$ of EMD is set to 0.0001 in this experiment, which is applied for early stopping. Generally, the detection accuracy gradually improves with the number of iterations until convergence. Moreover, the ablation study also reflects the convergence rate of the applied entropic mirror descent method. As shown in Table V, the applied EMD method exhibits reasonably high convergence rate, requiring only around 10 iterations to converge. For complex SGG tasks, such high convergence rate is very welcome.

Moreover, we also conduct another ablation study to explore the effect of the scoring function on the performance of the proposed CSL method. As demonstrated in Table VI, we evaluate the SGGDet performance of the proposed CSL method with different scoring function configurations. It can be seen that the detection accuracy steadily improves by incorporating more higher order potential terms into the scoring function. In particular, by inserting the global contextual information into the scoring function, one could greatly improve the resulting SGGDet performance.

B. Open Images V6

1) Benchmark: Open Images V6 [53] (301 object categories and 31 predicate categories) from Google is another popular
SGG benchmark, with a superior annotation quality. The dataset contains 126,368 training images, 1,813 validation images and 5,322 test images. In this experiment, we choose the same data processing protocols as in [22], [33], [53].

2) Evaluation Metrics: Based on the evaluation protocols in [22], [33], [53], we choose the following metrics for the Open Images V6 benchmark: the mean Recall@50 (mR@50), the regular Recall@50 ($R_{@50}$), the weighted mean AP of relationships ($wmAP_{rel}$) and the weighted mean AP of phrases ($wmAP_{phr}$). Like [22], [33], [53], the weight metric score is defined as: $score_{wtd} = 0.2 \times R_{@50} + 0.4 \times wmAP_{rel} + 0.4 \times wmAP_{phr}$.

3) Implementation Details: As in the case of the Visual Genome experiment, we employ ResNeXt-101-FPN [46] as the backbone and for the object detector we choose Faster RCNN [44]. Moreover, we freeze the parameters of the above models and apply the same bi-level data resampling strategy [14] as in the previous experiment. The batch size $bs$ is set to 12. Finally, we employ a two-layer MLP to construct the associated log scoring function and utilize an Adam optimizer with the learning rate of 0.0001.

4) Comparisons With State-of-the-Art Methods: In this experiment, for a fair comparison, several previous methods are re-implemented using the authors’ latest code. This is indicated by the † symbol. The results are presented in Table VII. It can be seen that the proposed CSL method achieves the state-of-the-art performance in all evaluation metrics on the Open Images V6 benchmark. Besides the regular $R_{@50}$ metric, it outperforms the previous methods by a large margin, especially in the more informative $mR_{@50}$ metric, which further verifies the effectiveness of the proposed method.

In fact, over the last several years, all the reported gains in performance for Open Images V6 benchmark are of similar magnitude, reflecting that the solution to the challenging SGG problem is continuously improving. More importantly, with the proposed constrained optimization paradigm, further performance gains will come from further advances in the development of scoring functions and their modelling. It is important that these advances are supported by the best possible formulation of the learning problem. Thus, the significance of the proposed constrained optimization paradigm is much greater than the marginal gains achieved in the experiments reported in the paper.

C. Visualization Results

In this section, we compare the visualization of the qualitative results of the ground-truth (GT), the baseline BGNN model, the proposed CSL method and the derived CSL+BA algorithm in the SGDet task, as demonstrated in Fig. 5. Specifically, compared with the state-of-the-art BGNN model, the proposed CSL method is capable of detecting the informative tail or body predicate categories rather than the common head predicate categories. For instance, in the middle image, the proposed CSL method could detect an additional informative triplet $<$sign hanging from building$>$. Besides, it can also detect the spatial informative predicates like on back of or behind. Moreover, with the balance adjustment strategy, the derived CSL+BA algorithm further improves its capability in detecting the more informative tail or body predicate categories. For example, in

| Method | mR@50 | R@50 | wmAP_rel | wmAP_phr | score_wtd |
|--------|-------|------|----------|----------|-----------|
| RelIN$^1$ [33] | 33.98 | 73.08 | 32.16 | 33.39 | 40.84 |
| RelIN$^*$ [33] | 37.20 | 75.34 | 34.21 | 44.18 | 41.97 |
| VCTree$^1$ [13] | 33.91 | 74.08 | 34.16 | 33.11 | 40.21 |
| G-GNN$^*$ [11] | 34.04 | 74.51 | 33.15 | 34.21 | 41.84 |
| Motifs$^*$ [29] | 32.68 | 71.63 | 29.91 | 31.59 | 38.93 |
| VCTree-TDE$^1$ [30] | 35.47 | 69.30 | 30.74 | 32.80 | 39.27 |
| GPS-Net$^*$ [22] | 35.26 | 74.81 | 32.85 | 33.98 | 41.69 |
| GPS-Net$^*$ [22] | 38.93 | 74.74 | 32.77 | 33.87 | 41.60 |
| BGNN [14] | 40.45 | 74.38 | 33.51 | 34.15 | 42.06 |
| CSL | 41.72 | 75.44 | 34.30 | 35.38 | 42.86 |

* Note: All the above methods use ResNeXt-101-FPN as the backbone. † means the re-sampling strategy [56] is applied in this method, and †† depicts the results reproduced using the latest code from the authors.
Fig. 5. Visualization of the qualitative results of the ground-truth (GT), the baseline BGNN model, the proposed CSL method and the derived CSL+BA algorithm in the SGDet task. The black, orange and green arrows represent the triplets with head predicate categories, the triplets with body or tail predicate categories and the triplets detected by models which are not included in GT, respectively. Compared with the baseline BGNN model, the scene graphs generated by the proposed CSL method and the derived CSL+BA algorithm are much closer to the ground-truth scene graph GT.

the top image, the derived CSL+BA method is able to detect an additional informative triplet < window part of door >, or even a new reasonable triplet < giraffe behind window > (which is not included in the ground-truth scene graph GT) for the bottom image. Compared with the baseline BGNN model, the scene graphs generated by the proposed IWSL method and the derived IWSL+BA algorithm are much closer to the ground-truth scene graph GT, which clearly demonstrates the superiority of the proposed methods.

V. CONCLUSION

In this paper, we proposed a novel constrained structure learning method for the SGG task, in which an explicit constrained variational inference objective is applied in the proposed MFVB framework. Unlike the previous SGG models, the proposed method formulates the SGG task as a more general constrained optimization problem, and investigates an alternative inference technique other than the ubiquitous message passing strategy. Specifically, a generic entropic mirror descent algorithm is applied to accomplish the constrained variational inference step, while the associated marginals in the proposed MFVB framework are inferred by a specific variable elimination technique. Finally, in extensive experiments on the popular Visual Genome and Open Images V6 benchmarks, we showed the proposed generic method outperforms the traditional message passing based SGG models. Currently, we solve the SGG tasks from a mean field variational Bayes perspective, which indicates no structural dependencies are considered within the variational distribution. Moreover, the proposed CSL method cannot deal with noisy label annotations, since we have not formulated the SGG task as a noisy label learning problem. Furthermore, the traditional two stages SGG paradigm, rather than the end-to-end one stage SGG paradigm, is adopted in the proposed CSL method, which may affect the performance for the Open Images V6 benchmark. Solving these issues would be our next targets.

REFERENCES

[1] P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6077–6086.
[2] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10685–10694.
[3] D. Teney, L. Liu, and A. van Den Hengel, “Graph-structured representations for visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1–9.
[4] J. Shi, H. Zhang, and J. Li, “Explainable and explicit visual reasoning over scene graphs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8376–8384.
[5] M. J. Wainwright et al., “Graphical models, exponential families, and variational inference,” Found. Trends Mach. Learn., vol. 1, no. 1/2, pp. 1–305, 2008.
[6] C. W. Fox and S. J. Roberts, “A tutorial on variational Bayesian inference,” Artif. Intell. Rev., vol. 38, no. 2, pp. 85–95, 2012.

[7] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5410–5419.

[8] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, “Scene graph generation from objects, phrases and region captions,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1261–1270.

[9] B. Dai, Y. Zhang, and D. Lin, “Detecting visual relationships with deep relational networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3076–3085.

[10] S. Woo, D. Kim, D. Cho, and I. S. Kweon, “LinkNet: Relational embedding for scene graph,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 560–570.

[11] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph R-CNN for scene graph generation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 670–685.

[12] W. Wang, R. Wang, S. Shan, and X. Chen, “Exploring context and visual pattern of relationship for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8188–8197.

[13] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, “Learning to compose dynamic tree structures for visual contexts,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6619–6628.

[14] R. Li, S. Zhang, B. Wan, and X. He, “Bipartite graph network with adaptive message passing for unbiased scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11109–11119.

[15] D. Liu, M. Bober, and J. Kittler, “Visual semantic information pursuit: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 4, pp. 1404–1422, Apr. 2021.

[16] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” IEEE Trans. Neural Netw., vol. 20, no. 1, pp. 61–80, Jan. 2009.

[17] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in Proc. Int. Conf. Mach. Learn., PMLR, 2017, pp. 1263–1272.

[18] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7794–7803.

[19] J. Zhou et al., “Graph neural networks: A review of methods and applications,” AI Open, vol. 1, pp. 57–81, 2020.

[20] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, “Factorizable net: An efficient subgraph-based framework for scene graph generation,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 335–351.

[21] T. Chen, W. Yu, R. Chen, and L. Lin, “Knowledge-embedded routing network for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6163–6171.

[22] X. Lin, C. Ding, J. Zeng, and D. Tao, “GPS-net: Graph property sensing network for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3746–3753.

[23] C. Sutton and M. McCallum, “An introduction to conditional random fields for relational learning,” Introduction Statist. Relational Learn., vol. 2, pp. 93–128, 2006.

[24] W. Ping, Q. Liu, and A. Ihler, “Marginal structured SVM with hidden variables,” in Proc. Int. Conf. Mach. Learn., 2014, pp. 190–198.

[25] H. Liu, N. Yan, M. Mortazavi, and B. Hanu, “Fully convolutional scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11546–11556.

[26] Y. Cong, M. Y. Yang, and B. Rosenhahn, “ReTrKR: Relation transformer for scene graph generation,” 2022, arXiv:2201.11460.

[27] Y. Teng and L. Wang, “Structured sparse R-CNN for direct scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4010–4019.

[28] R. Li, S. Zhang, and X. He, “SGTR: End-to-end scene graph generation with transformer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 19486–19496.

[29] Z. Rizzelli, M. Yatskar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5831–5840.

[30] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, “Unbiased scene graph generation from biased training,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3716–3725.

[31] G. Raskutti and S. Mukherjee, “The information geometry of mirror descent,” IEEE Trans. Inf. Theory, vol. 61, no. 3, pp. 1451–1457, Mar. 2015.

[32] Y. Zhu and S. Jiang, “Deep structured learning for visual relationship detection,” in Proc. AAAI Conf. Artif. Intell., 2018, Art. no. 934.
Daqi Liu is a research fellow with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. His research interests include machine learning, computer vision, pattern recognition, and bio-inspired computational models. He has published several scientific papers in top-ranked journals, including *IEEE transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Cybernetics* etc.

Miroslaw Bober (Member, IEEE) received the MSc and PhD degrees from the University of Surrey, Guildford, U.K., in 1991 and 1995, respectively. He is a professor of video processing with the University of Surrey, Guildford, U.K. Between 1997 and 2011, he headed the Mitsubishi Electric Corporate R&D Center Europe (MERCE), Livingston, U.K. He has been actively involved in the development of MPEG standards for more than 20 years, chairing the MPEG-7, CDVS, and CVDA groups. He is an inventor of more than 70 patents and several of his inventions are deployed in consumer and professional products. He has authored or coauthored more than 80 refereed publications, including three books and book chapters. His research interests include various aspects of computer vision and machine intelligence, with recent focus on image/video database retrieval and data mining.

Josef Kittler (Life Member, IEEE) received the BA, PhD, and DSc degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished professor of machine intelligence with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook *Pattern Recognition: A Statistical Approach* and more than 700 scientific papers. His publications have been cited more than 60,000 times (Google Scholar). He was a series editor of Springer Lecture Notes on Computer Science. He currently serves on the editorial boards of *Pattern Recognition Letters, Pattern Recognition, and Springer Nature: Computer Science*. He also served as a member of the editorial board of *IEEE Transactions on Pattern Analysis and Machine Intelligence* during 1982–1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982–2005, President of the IAPR during 1994–1996.