This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3243306

Neural Belief Propagation for Scene Graph Generation

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

Abstract—Scene graph generation aims to interpret an input image by explicitly modelling the objects contained therein and their relationships. In existing methods the problem is predominantly solved by message passing neural network models. Unfortunately, in such models, the variational distributions generally ignore the structural dependencies among the output variables, and most of the scoring functions only consider pairwise dependencies. This can lead to inconsistent interpretations. In this paper, we propose a novel neural belief propagation method seeking to replace the traditional mean field approximation with a structural Bethe approximation. To find a better bias-variance trade-off, higher-order dependencies among three or more output variables are also incorporated into the relevant scoring function. The proposed method achieves the state-of-the-art performance on various popular scene graph generation benchmarks.

Index Terms—Scene Graph Generation, Message Passing, Variational Approximation, Graph Neural Networks, Belief Propagation.

1 Introduction

Given an input image, scene graph generation (SGG) seeks to explicitly model the objects appearing in the scene and their relationships. It is a structured prediction task, in which the output interpretations (or labels) are usually conditionally dependent. In current SGG models [1], [2], [3], [4], [5], [6], [7], [8], the discriminative undirected probabilistic graphical models like conditional random fields (CRFs) [9], [10] are often applied to model the above conditional dependencies using the relevant scoring functions. Due to the combinatorial nature of the structured outputs in SGG applications, it is generally computationally intractable to compute the underlying posterior directly.

To this end, variational Bayes (VB) modelling [11], [12] is generally employed for the SGG tasks, in which the variaitonal inference step aims to pursue the optimum interpretations using a maximum aposteriori (MAP) inference, while the variational learning step seeks to fit the model posterior with the ground-truth training samples via a classical cross entropy loss. For tractability, the variational distribution in current SGG models is generally assumed to be fully decomposable, and the resulting VB is also known as mean field variational Bayes (MFVB) [11], [12].

Specifically, message passing neural networks (MPNNs) [4], [5], [6], [7], [8] are generally employed to model the above MFVB framework aiming to leverage the inference capability of the MFVB as well as the feature representation learning ability of the deep learning models. The resulting MPNN-based MFVB formulation has became the de facto methodology for the current SGG models, in which two fundamental modules are required, namely, visual perception and visual context reasoning [13]. The former extracts a set of region proposals within the input image while the latter infers the optimum instance/relationship interpretations for those region proposals.

• *The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K.*
E-mail: {daqi.liu, m.bober, j.kittler}@surrey.ac.uk

However, the above MPNN-based MFVB formulation suffers from two main drawbacks: 1) Most of the applied scoring functions in the current SGG models only consider pairwise dependencies. The higher-order dependencies among three or more output variables are largely ignored. Such higher-order dependencies play a vital role in generating consistent interpretations since they constrain the possibilities of certain interpretation combinations. For instance, $< man\ play\ football >$ is more likely to occur than $< plant\ play\ football >$. Without the relevant third-order conditional dependencies, one can not guarantee the above preference or tendency; 2) The variational distribution used to approximate the model posterior generally assumes the output variables are totally independent without any structural dependencies. Such factorized distributions enable efficient variational inference but they sacrifice the accuracy [14]. In the true posterior, many latent variables are dependent and the mean field approximation, by construction, fails to capture this dependency [15].

To address the above issues, inspired by the recently proposed factor graph neural network (FGNN) model [16] tailored for point cloud segmentation tasks, we propose a novel neural belief propagation (NBP) paradigm aiming to replace the previous mean field approximation [11], [12] with a structural Bethe approximation [17], [18]. This is because the Bethe approximation, compared with the naive mean field approximation, has the potential to provide a better model of the log evidence [19]. To capture higher-order dependencies other than the common pairwise dependencies, a new scoring function is defined in this paper, which enable us to find a better bias-variance trade-off [20]. More importantly, the proposed NBP paradigm, rather than applying the max-product rule as in the FGNN model, simulates a sum-product strategy in order to avoid underestimating the model posterior [21]. Unlike the previous ubiquitous MPNN-based MFVB SGG models, the proposed NBP paradigm combines the inference ability of the Bethe approximation strategy and the feature representation learn-

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3243306

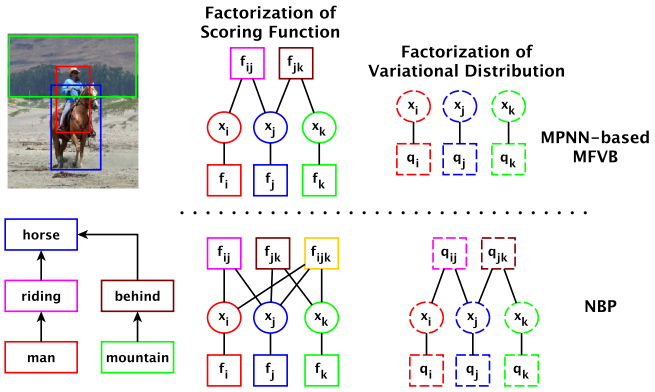


Fig. 1: Comparison between the traditional MPNN-based MFVB paradigm and the proposed NBP paradigm. In SGG, given an input image I, a factor graph representing the factorization of a scoring function is constructed, in which x and f represent the object and factor (relationship) vertices, respectively. A variational distribution $q(x)$ is often used to approximate the computationally intractable model posterior $p(x|I)$ derived from the relevant scoring function. Unlike MPNN-based MFVB, NBP includes certain higher-order dependencies (f_{ijk}) into the scoring function aiming to find a better bias-variance trade-off, and incorporates relevant pairwise dependencies (q_{ij} and q_{jk}) into the variational distribution seeking to estimate the model posterior better.

ing capability of the deep neural networks. The specific comparison between the traditional MPNN-based MFVB paradigm and the proposed NBP paradigm is demonstrated in Fig.1. The proposed generic NBP paradigm achieves the state-of-the-art performance on two popular SGG benchmarks: Visual Genome and Open Images V6.

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

1) A novel NBP paradigm is proposed to replace the traditional MPNN-based MFVB paradigm aiming to solve the two main drawbacks mentioned above.
2) To incorporate the pairwise dependencies into the variational distribution, a novel Factor-to-Variable and Variable-to-Factor MPNN framework is proposed in the NBP paradigm. Unlike FGNN model which seeks to address a MAP inference task, it aims to solve a relevant marginal inference task.
3) Two novel node adjacency sets are designed in the NBP paradigm to define the edges of the applied factor graph, which aims to include certain higher-order dependencies into the relevant scoring function.
4) To our knowledge, we are the first one to apply such NBP paradigm to accomplish the complex SGG tasks.

This paper is organized as follows: Section 2 and Section 3 present the related work and the proposed neural belief propagation method, respectively. The experimental results and the associated analysis are elaborated in Section 4. Finally, the conclusions are drawn in Section 5.

2 Related Work

As structured prediction tasks, current SGG applications often aim to achieve two main objectives: extracting informative feature representations and implementing unbiased long-tailed recognition. They could facilitate the downstream computer vision tasks like image captioning [22], [23], [24] or visual question answering [25], [26], [27].

2.1 Extracting Informative Representations

Current SGG models often rely on a MPNN-based MFVB paradigm to learn discriminative representation for node and edge prediction. They mainly focus on graph structure design and contextual feature fusion strategies via message passing mechanisms.

To achieve context modelling, various graph structures have been proposed in recent years. A popular choice is to apply a sequential model [4] or a fully-connected graph to model the context [5], [28]. Another choice is to investigate the sparse graph structures [24] by associating the downstream tasks or trimming the relationship proposals based on category or geometry information of subject-object pairs.

To incorporate the context information into the existing SGG models, different message passing mechanisms have been explored. Some of them apply message passing between the entities proposals [7], [8], [29] while others aggregate the contextual information between the entities and predicates [6].

2.2 Long-tailed Recognition

To alleviate the ubiquitous biased relationship prediction problem caused by the long-tail data distribution, three directions are investigated in literature: data re-sampling, cost-sensitive losses and transfer learning.

Specifically, data re-sampling tries to over-sample underrepresented (tail) classes and under-sample populated (head) classes, aiming to generate a more uniform training distribution. Specifically, oversampling is often achieved by duplicating samples or by synthesizing data. The representative data re-sampling strategies include dataset resampling [30], [31], [32], instance-level resampling [33], [34] and a mixed bi-level data resampling [35]. Based on the class frequency or difficulty, different costs are assigned in the cost-sensitive losses [36], [37] to the incorrect prediction of different samples. For classes with fewer samples, one should assign higher weights or enforce larger margins. Weights can be estimated by the inverse class frequency or meta-learning. Transfer learning methods [38], [39], [40] aim to transfer information from head to tail classes. Recently, [41] formulates the SGG task as a causal model and present an unbiased learning method based on causal inference.

Unlike the above mainstream SGG models which naively assume the output variables in the variational distribution are independent, the proposed NBP method incorporates the structural pairwise dependencies into the corresponding variational distribution. The factorized variational distributions applied in the mainstream SGG models enable efficient variational inference but they sacrifice the accuracy [14]. Motivated by [16], [42], [43], the approach aims to solve the SGG tasks by simulating a belief propagation

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3243306

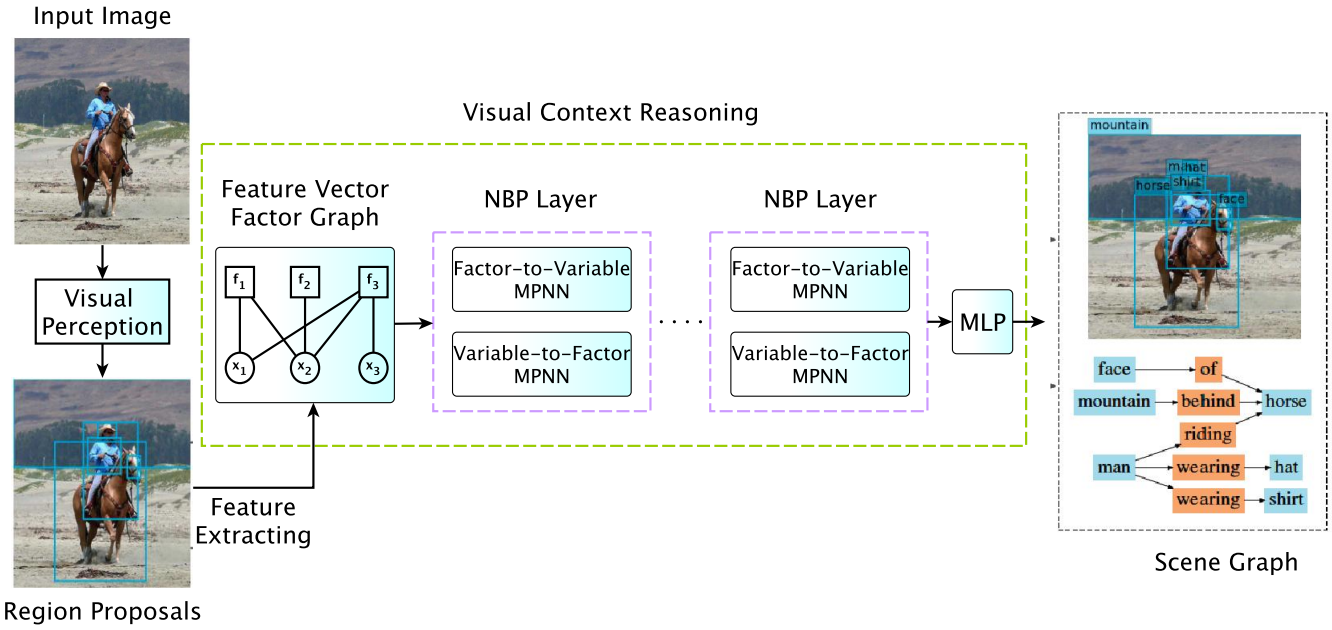


Fig. 2: Overview of the proposed neural belief propagation (NBP) method, in which the green dash line represents the proposed visual context reasoning module. Given an input image, visual perception module detects a set of region proposals. A feature vector factor graph $G = (X, F, E)$ is constructed based on the above region proposals, in which the vertex variable set X, the factor set F and the related edge set E are all associated with the corresponding feature vectors. With such a feature vector factor graph, one can infer the resulting marginals via stacking various NBP layers plus a final MLP. Each of the NBP layers consists of two message passing neural network (MPNN) models: Factor-to-Variable MPNN and Variable-to-Factor MPNN, which essentially correspond to the two types of messages within the classical belief propagation method. Given the resulting marginals of the factor graph G, one can easily compute the optimum interpretations (for related vertex variables and pairwise factors, which corresponds to instances and predicates in a scene graph) via simple $argmax$ operations. A cross-entropy loss is applied to train the proposed NBP method.

methodology rather than the classical MFVB framework. To our knowledge, the proposed NBP method is the first one to use the neural belief propagation type methodology to solve the complex SGG tasks. Specifically, to find a better bias-variance trade-off [20], a new scoring function with higher-order dependencies is proposed and the structure of the related factor graph is defined by two novel node adjacency sets. Compared with the ubiquitous mean field approximation applied in the mainstream SGG models, the structural Bethe approximation employed in the proposed NBP method is a better estimation of the log evidence [19]. Besides, the mean field approximation often underestimates the underlying model posterior [14].

3 Proposed Methodology

In this section, we first formulate the SGG problem. This will be followed by the development of a novel scoring function and a discussion of the proposed neural belief propagation method. Fig.2 shows an overview of the proposed neural belief propagation method, which is described in detail in the figure caption.

3.1 Problem Formulation

As a structured prediction task, given an input image, SGG aims to model the potential objects as well as their relationships via certain inference strategies. The scene graph consists of a set of intertwined semantic triplet structures, in which each triplet includes three components: a subject, a predicate and an object. In the current SGG settings, one only focuses on inferring the pairwise relationships, where the relationship between two interacting instances (subject and object) in an input image is termed as a predicate. Such structured prediction task can be naturally modelled by a discriminative undirected probabilistic graphical model, i.e. Conditional Random Field.

Generally, the above undirected graphical model can be further transformed into a factor graph [11], [44]. It is a bipartite probabilistic graphical model, which aims to model the dependencies among the random variables via factorizing a corresponding scoring function:

$$S(x) = \prod_{r \in R} f_r(x_r) \tag{1}$$

where the dependencies among each subset of variables (clique) x_r are modeled by a corresponding non-negative factor function f_r. Fig.3 demonstrates a simple factor graph, in which factor f_1 depends on subset $\{x_1\}$, f_2 depends on subset $\{x_1, x_3\}$ and f_3 depends on subset $\{x_2, x_3\}$. Here, f_1 corresponds to a 1-vertex clique, while f_2 and f_3 have 2-vertex cliques. With the above scoring function $S(x)$, one can compute the corresponding probability distribution $p(x) = \prod_{r \in R} f_r(x_r)/Z$, where Z is the associated partition function.

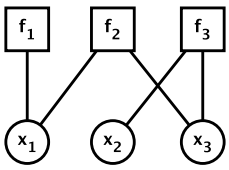


Fig. 3: An example factor graph. f_1 corresponds to 1-vertex (subset with only one vertex variable $\{x_1\}$) clique, while f_2 and f_3 have 2-vertex (subsets with two vertex variables $\{x_1, x_3\}$ and $\{x_2, x_3\}$, respectively) cliques.

Given the above factor graph, a classical belief propagation method [17], [18] is commonly applied to accomplish the relevant inference tasks. As a message passing method, belief propagation performs inference on graphical models by locally marginalizing over random variables, which is also known as the sum-product algorithm. It can efficiently compute the associated marginals via exploring the unique structure of the applied factor graph. Specifically, it works by sending real-valued functions called messages along the associated edges. With such messages, the nodes can exchange their beliefs about other nodes and thus transporting the associated probabilities. Based on whether the node receiving the message is a variable node or factor node, there are two types of messages: Variable-to-Factor message and Factor-to-Variable message. Specifically, Variable-to-Factor message $\mu_{x_i \to f_a}$ is the product of the messages from all other neighboring factor nodes $N(x_i)$ except f_a:

$$\mu_{x_i \to f_a} = \prod_{a^* \in N(x_i) \setminus f_a} \mu_{f_{a^*} \to x_i}(x_i) \tag{2}$$

while Factor-to-Variable message is the product of the factor with messages from all other nodes, marginalized over all variables x_a except x_j:

$$\mu_{f_a \to x_j} = \sum_{x_a \setminus x_j} f_a(x_a) \prod_{i \in N(f_a) \setminus j} \mu_{x_i \to f_a}(x_i) \tag{3}$$

where, after recursively running the above two steps until convergence, the associated marginal $p(x_j)$ can be computed as:

$$p(x_j) \propto \prod_{a \in N(x_j)} \mu_{f_a \to x_j}(x_j) \tag{4}$$

Specifically, suppose we can model a target SGG task using a factor graph $G = (X, F, E)$, in which X represents a variable vertex set including all the potential instances $(x_1, x_2, ..., x_u)$ detected by the associated visual perception module, F is a corresponding factor set containing all the potential relationships $(f_1, f_2, ..., f_w)$ among the instances, E is an edge set in which the edge e_{ij} only exists if the factor f_j is connected to the variable vertex x_i. Here, u and w represent the number of instances and relationships, respectively. The corresponding scoring function for $G = (X, F, E)$ can be described as follows:

$$S(I, x) = S(I, x_1, x_2, ..., x_u) = \prod_{j}^{w} f_j(I, C_j) \tag{5}$$

where $C_j \subseteq \{x_1, x_2, ..., x_u\}$ is a potential clique and f_j represents a non-negative factor function which characterizes the dependencies among the vertex variables within C_j.

With the above scoring function, one can directly compute the model posterior $p(x|I)$ as:

$$p(x|I) = \frac{S(I, x)}{\sum_x S(I, x)} = \frac{S(I, x)}{S(I)} \tag{6}$$

where $S(I)$ is the partition function. Due to the exponential structural outputs in SGG tasks, such model posterior $p(x|I)$ is generally computationally intractable. Therefore, variational inference (VI) technique is often used to approximate $p(x|I)$ with a computationally tractable variational distribution $q(x)$. For tractability, $q(x)$ assumes the output variables are fully independent, and it can decomposed as follows:

$$q(x) = \prod_{i=1}^{u} q_i(x_i) \tag{7}$$

where $q_i(x_i)$ represents the local variational approximation of the i-th output variable x_i. Such VI is also known as mean field variational inference (MFVI), and the relevant VB framework is often called mean field variational Bayes (MFVB) [11], [12]. Specifically, current SGG models are generally formulated as a MPNN-based MFVB framework, in which two fundamental modules are often required, namely, visual perception and visual context reasoning [13]. The former extracts a set of region proposals within the input image while the latter infers the optimum instance/relationship interpretations for those region proposals.

Unlike the above traditional MPNN-based MFVB framework, in this paper, a neural belief propagation methodology is employed to solve the SGG tasks, which leverages both the inference capability of the belief propagation method and the feature representation learning ability of the deep neural networks. Given an instance interpretation set $\mathcal{C}$ and a relationship interpretation set $\mathcal{R}$, an SGG task can be modelled as a corresponding genetic factor graph $G = (X, F, E)$, in which one needs to infer each marginal distribution corresponding to each vertex variable $x_i \in \mathcal{C}, i = 1, 2, ..., u$ within X as well as each factor $f_j \in \mathcal{R}, j = 1, 2, ..., w$ within F. In current SGG settings, only pairwise dependencies (2-vertex cliques) are required to be scored. Specifically, to efficiently infer the relevant marginal distributions from the above factor graph $G = (X, F, E)$, a neural belief propagation method implementing the classical sum-product rule is employed. With the above marginal distributions, one can easily compute the optimum interpretations x^* via additional $argmax$ operations.

3.2 Proposed Scoring Function

In current SGG models, the scoring function $S(I, x)$ only incorporates two types of cliques: 1-vertex cliques $C = \{x_i\}, i = 1, 2, ..., u$ and 2-vertex cliques $C = \{x_i, x_j\}, j \in N(i)$, where j is the vertex around the target vertex i. The former measures the similarity between I and each potential variable vertex x_i while the latter characterizes

the dependency between two interacting variable vertexes $\{x_i, x_j\}$.

$$S(I,x) = \prod_{i}^{u}[f_i(I,x_i) \prod_{j \in N(i)} f_{ij}(x_j,x_i)] \tag{8}$$

Such scoring function formulation only considers the pairwise dependencies among the vertex variables, which may underestimate the ground-truth posterior $p_r(x|I)$. In the ground-truth posterior $p_r(x|I)$, higher-order dependency among the latent variables commonly exist, yet the scoring function applied in the current SGG models, by construction, fails to capture this higher-order dependency. As a result, the model posterior $p(x|I)$ produced by such a scoring function is not a tight approximation of the ground-truth posterior $p_r(x|I)$. The consequence of the approximation error is a model bias.

To lower the above model bias and find a better bias-variance trade-off, in this paper, we propose a novel scoring function formulation, which incorporates multi-vertex (higher than 2-vertex) cliques into the associated scoring function:

$$S(I,x) = \prod_{i}^{u}[f_i(I,x_i) \prod_{j \in N(i)} f_{ij}(x_j,x_i)] \prod_{h} f_h(x_h) \tag{9}$$

where $h \subseteq \{x_1, x_2, ..., x_u\}$ is a multi-vertex clique with f_h as its corresponding factor function. In the proposed scoring function formulation, for an input image I with u detected vertex variables $\{x_1, x_2, ..., x_u\}$, the multi-vertex clique $h = \{x_1, x_2, ..., x_u\}$ includes all the potential detected vertex variables.

For feasibility and tractability, we choose the current higher-order clique (involving all the detected instances in an input image) rather than other types of higher-order cliques (with 3-vertex or 4-vertex). Specifically, with our proposed higher-order clique, one can easily define the relevant clique structure by including all the detected instances in an input image, and only one corresponding higher-order potential term is required to incorporate into the scoring function. In contrast, with other types of higher-order cliques, we need to add combinatorial higher-order potential terms into the scoring function, and improving the complexity of a model (e.g. incorporating more high-order potential terms into the scoring function) would substantially increase the computation burden. More importantly, it is hard to define the structures of other higher-order cliques since only the structures of the pairwise cliques (pairwise relationships) are well-defined in the current SGG settings.

The above formulation considers the global contextual information via incorporating the above multi-vertex cliques into the target scoring function. Essentially, it lowers the model bias by aiming to pursue a better bias-variance trade-off. Traditionally, improving the complexity of a model would undoubtedly increase the associated computational burden. To this end, a novel neural belief propagation method is introduced in the following subsection.

3.3 Neural Belief Propagation

To efficiently infer the associated marginal distributions for the above complex scoring function, in this section, we propose a novel neural belief propagation (NBP) method, which combines the powers of both classical belief propagation algorithm and the modern message passing neural network structures. The proposed NBP method extends the current message passing neural network so that it can break the previous universal yet naive independence assumption. To better illustrate the proposed method, we first present a generic message passing neural network framework, followed by the introduction of a specific FGNN structure. Finally, we explain how we build the proposed NBP method based on such an FGNN structure.

3.3.1 Generic Message Passing Neural Network

Following [45], in this section, we present a generic message passing neural network (MPNN) structure for the current SGG models. With such generic framework, one can easily define a new graph neural network model by modifying the related message passing operations.

Specifically, given a graph $G = (X, E)$ with a set of nodes X and a set of pairwise edges E, suppose each node x_i is associated with a feature representation v_i and each possible pairwise edge e_{ij} is associated with a feature representation t_{ij} (where $j \in N(i)$ is a neighbouring node to i), the above generic message passing neural network can be described as:

$$m_i = \sum_{j \in N(i)} M(v_i, v_j, t_{ij}),\ \hat{v}_i = U(v_i, m_i) \tag{10}$$

where m_i represents the intermediate message obtained from a relevant neural network M, $\hat{v}_i$ is the updated feature representation of the node x_i by feeding v_i and m_i into a corresponding neural network U. The summation aggregator in the above equation can be replaced by other operations. Such a generic framework can be generally applied to any graph with only pairwise edges.

3.3.2 Factor Graph Neural Network

To extend the above formulation to the generic factor graphs with extra edges, other than the pairwise ones, a factor graph neural network (FGNN) structure [16] has recently been proposed. Specifically, a FGNN layer follows a unique MPNN structure, which can be denoted as follows:

$$\hat{v}_i = \max_{j \in N(i)} Q(t_{ij}) M(v_i, v_j) \tag{11}$$

where the functions Q and M are implemented by neural networks, in which $Q : \mathbb{R}^{d_{in}} \rightarrow \mathbb{R}^{m \times n}$ transforms the input $t_{ij} \in \mathbb{R}^{d_{in}}$ into an $m \times n$ weight matrix, while $M : \mathbb{R}^{d_{in}+d_{in}} \rightarrow \mathbb{R}^{n}$ maps the concatenated feature vector $[v_i, v_j] \in \mathbb{R}^{d_{in}+d_{in}}$ into a length-n feature vector. Here, d_{in} represents the input feature dimension, m denotes the dimension of the output feature vector and n is a hyper parameter for Q net. Correspondingly, a new length-m feature vector $\hat{v}_i$ is produced after the above matrix multiplication and aggregation.

With the above unique MPNN structure, FGNN encodes the higher order features via incorporating extra factor nodes. Consider a factor graph $G = (X, F, E)$. Suppose each vertex variable $x_i \in X$ is associated with a feature variable v_i, each factor $f_j \in F$ is associated with a factor feature g_j, and each edge $e_{ij}, j \in N(i)$ connecting x_i and

f_j is associated with an edge representation t_{ij}. An FGNN layer consists of two important modules: Variable-to-Factor MPNN and Factor-to-Variable MPNN, which are defined as follows:

$$\begin{aligned} \hat{g}_j &= \max_{i \in N(j)} Q(t_{ij}|\phi_{VF}) M([v_i, g_j]|\psi_{VF}) \\ \hat{v}_i &= \max_{j \in N(i)} Q(t_{ij}|\phi_{FV}) M([v_i, g_j]|\psi_{FV}) \end{aligned} \tag{12}$$

where ϕ and ψ are used to parameterize the neural networks Q and M, respectively. As shown in the above equation, those two MPNN modules have similar structures but different parameters. More importantly, one can simulate k max-product iterations via stacking k FGNN layers, plus a linear layer at the end. In other words, with the above FGNN structure, one can perform inference over the associated factor graph akin to the classical max-product method.

3.3.3 Neural Belief Propagation Method

To extend the above FGNN model to efficiently solve the complex SGG tasks, we propose a novel neural belief propagation (NBP) method in this paper. In contrast to the FGNN model which seeks to accomplish a MAP inference task, the proposed NBP method aims to solve a relevant marginal inference task. In other words, it simulates a sum-product rule rather than the max-product strategy aiming to avoid underestimating the model posterior [21].

Specifically, given an input image I, SGG aims to infer the optimum interpretations x^* as:

$$x^* = \arg\max_x p(x|I) = \arg\max_x S(I, x) \tag{13}$$

where $p(x|I)$ represents the model posterior and $S(I, x)$ denotes the associated scoring function. The above inference task is essentially formulated as a maximum aposteriori estimation problem (MAP), which is generally NP-hard to solve for structured prediction tasks like SGG. Following [21], such NP-hard MAP inference can be formulated as an integer linear program and further transformed into a relevant relaxed linear program. More importantly, one can unify the above MAP inference with the related marginal inference as follows:

$$q^* = \arg\max_q \mathbb{E}_{q(x)} S(I, x) + T\mathbb{H}(q(x)) \tag{14}$$

where $q(x)$ is a variational distribution and $\mathbb{H}(q(x))$ is its entropy. T is a temperature parameter where $T = 1$ for marginal inference and $T = 0$ for MAP inference. On the right hand side of Equation (14), the first expectation term leads to a unimodal variational distribution since it prefers $q(x)$ to place its mass on the MAP estimate, while the second entropy term produces a multimodal variational distribution since it encourages $q(x)$ to be diffuse. Compared with the MAP inference, the marginal inference could find a trade-off between the above two scenarios and thus potentially avoid underestimating the complex model posterior $p(x|I)$. To this end, in this paper, T is often set to 1, and we prefer to first infer the marginals and then compute the optimum interpretations x^* via additional $argmax$ operations. Empirically, we find the above solution often produces a better performance than the exact MAP inference route. In particular, two fundamental modules are often required in current SGG models, namely, visual perception and visual context reasoning. The former seeks to locate and instantiate the objects and predicates within the input image, while the latter aims to infer their consistent interpretations.

Given an input image I, visual perception module produces a set of object region proposals $b_i^o \in \mathbb{R}^4, i = 1, ..., u$, as well as a set of predicate region proposals $b_j^p \in \mathbb{R}^4, j = 1, ..., w$, where u and w represent the number of instances and predicates detected in the input image, respectively. Correspondingly, one could extract the relevant vertex feature set $v_i \in \mathbb{R}^d, i = 1, ..., u$ and the related pairwise factor feature set $g_j \in \mathbb{R}^d, j = 1, ..., w$ by applying a ROI pooling on the feature maps obtained from the visual perception module. Given a set of object classes $\mathcal{C}$ and a set of relationship categories $\mathcal{R}$, a visual context reasoning module aims to infer the resulting instance and predicate interpretation sets $x_i \in \mathcal{C}, i = 1, ..., u$ and $f_j \in \mathcal{R}, j = 1, ..., w$ based on the above latent feature representation sets. Besides, in this paper, a higher-order region proposal $b^h \in \mathbb{R}^4$ is further obtained by computing the bounding box of the union of the bounding boxes of all the instances within the input image. With b^h, one could also extract a corresponding higher-order factor feature g_h. We argue, that it is beneficial to incorporate the higher-order factor node f_h into the applied scoring function, even though it is not required to be classified in current SGG tasks. Up until now, we have already associated the corresponding feature representations for both vertex variables X and factor variables F.

Furthermore, to complete the feature representation association task for the factor graph $G = (X, F, E)$, one needs to first define the relevant edges E and then associate feature representations for E. In particular, to specify the edges between the vertex variables X and the factor variables F, two novel node adjacency sets are proposed, based on the proposed scoring function, namely, pairwise node adjacency set and higher-order node adjacency set. The former aims to capture two types of pairwise dependency: 1) dependency between an instance and a predicate (e.g. man and $walking\ on$); 2) dependency between an instance and another instance (e.g. man and $street$), while the latter seeks to capture the higher-order dependency among all the involving nodes within the input image (e.g. man, $street$, $walking\ on$).

Moreover, to leverage the inference capability of the sum-product rule and the feature representation learning capability of the message passing neural networks, according to the nature of SGG tasks, novel Variable-to-Factor and Factor-to-Variable MPNN structures are proposed to construct an NBP layer. Suppose we have c NBP layers, a feature dimension list $(d_1, d_2, ..., d_c)$ is predefined to specify the dimensions of the output feature representations of those NBP layers. Within the l-th layer, given the vertex feature $v_i^l \in \mathbb{R}^{d_l}$ and the factor feature $g_j^l \in \mathbb{R}^{d_l}$, the proposed Variable-to-Factor and Factor-to-Variable MPNN structures are defined as follows:

$$\begin{aligned} g_j^{l+1} &= H(\sum_{i \in N(j)} Q(T([v_i^l, g_j^l]) \mid \phi_{VF}) M([v_i^l, g_j^l] \mid \psi_{VF})) \\ v_i^{l+1} &= H(\sum_{j \in N(i)} Q(T([v_i^l, g_j^l]) \mid \phi_{FV}) M([v_i^l, g_j^l] \mid \psi_{FV})) \end{aligned} \tag{15}$$

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3243306

where $v_i^{l+1} \in \mathbb{R}^{d_{l+1}}$ and $g_j^{l+1} \in \mathbb{R}^{d_{l+1}}$ are the updated vertex and factor feature representations, to be used as the input vertex and factor feature representations in the next $(l+1)$-th NBP layer. $H : \mathbb{R}^m \rightarrow \mathbb{R}^{d_{l+1}}$ is a neural network to map the intermediate length-m feature representation to the final length-d_{l+1} feature representation. N represents the above node adjacency sets, which are used to define the edges of the applied factor graph. $T : \mathbb{R}^{d_l+d_l} \rightarrow \mathbb{R}^{d_l}$ is a neural network aiming to produce a feature representation for each possible edge. ϕ and ψ are employed to parameterize the associated neural networks $Q : \mathbb{R}^{d_l} \rightarrow \mathbb{R}^{m\times n}$ and $M : \mathbb{R}^{d_l+d_l} \rightarrow \mathbb{R}^n$, respectively. As shown in the above equation, the applied Variable-to-Factor and Factor-to-Variable models have the same MPNN structures, but with different parameterizations (depicted using the subscripts VF and FV).

In particular, within the l-th NBP layer, for each potential edge within the node adjacency sets N, we employ a neural network $T : \mathbb{R}^{d_l+d_l} \rightarrow \mathbb{R}^{d_l}$ to map the concatenation of the vertex feature $v_i^l \in \mathbb{R}^{d_l}$ and the factor feature $g_j^l \in \mathbb{R}^{d_l}$ into an edge feature representation, which is then transformed by an $m \times n$ matrix via another neural network $Q : \mathbb{R}^{d_l} \rightarrow \mathbb{R}^{m\times n}$. The feature representation concatenation $[v_i^l, g_j^l]$ is subsequently mapped into a length-n feature representation via a neural network $M : \mathbb{R}^{d_l+d_l} \rightarrow \mathbb{R}^n$. Correspondingly, a new length-m feature representation is obtained via a relevant matrix multiplication for each potential edge. Furthermore, a mean aggreagator is applied for all related edges to produce the resulting m dimensional feature representation. Finally, a neural network $H : \mathbb{R}^m \rightarrow \mathbb{R}^{d_{l+1}}$ is employed to map the above m dimensional feature to the final length-d_{l+1} feature representation, which is used as the input to the next $(l+1)$-th NBP layer. In this paper, we use a multi-layer perceptron (MLP) to implement the above associated neural networks.

Finally, with the above NBP layers, for each vertex variable (instance) and each pairwise factor variable (predicate) in $G = (X, F, E)$, an $m\times 1$ feature vector is produced, which is further mapped into a corresponding logit via an associated MLP. A cross entropy loss is employed to train the above NBP method. As a result, the proposed NBP method extends the current message passing neural networks, in which the pairwise dependencies are incorporated into the associated variational approximation. In other words, one can apply a structural Bethe approximation [17], [18] to replace the previous ubiquitous mean field approximation. By virtue of the proposed NBP method, a tighter variational approximation is obtained for the underlying model posterior $p(x|I)$ [17], [18].

4 Experiments

To validate the proposed neural belief propagation method, two popular scene graph generation benchmarks - Visual Genome [46] and Open Images V6 [47] - are utilized in this section. For each benchmark, we first introduce the experimental configuration, followed by the comparisons with the state-of-the-art methods. The ablation study and the visualization results are also included in the last two subsections.

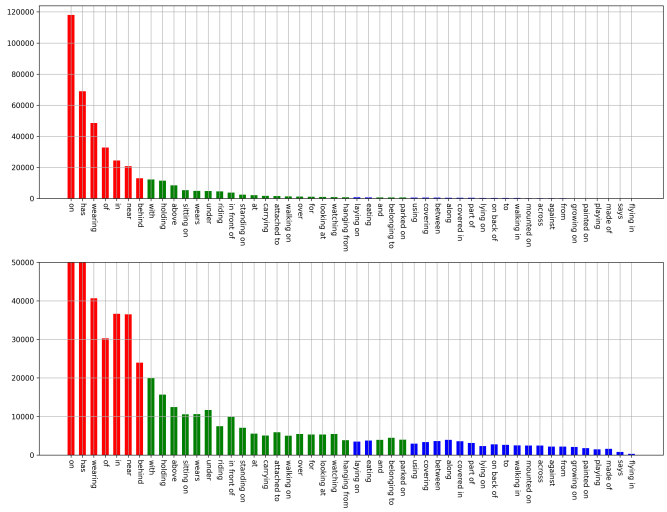


Fig. 4: Three disjoint category groups in Visual Genome training split (follows a long-tail data distribution as demonstrated above): $head$ (red bars), $body$ (green bars) and $tail$ (blue bars). y axis represents the number of samples.

4.1 Visual Genome

4.1.1 Experimental Configuration

Benchmark: As the most popular SGG benchmark, Visual Genome [46] consists of 108,077 images with an average of 38 objects and 22 relationships per image. Following the data split protocol in [50], in this experiment, we choose the most frequent 150 object classes and 50 predicate classes. Specifically, we split Visual Genome benchmark into two sets: a training set (70%) and a test set (30%). An evaluation set ($5k$) is further extracted from the training set for model validation. To investigate the biased relationship prediction problem caused by long-tail data distribution, according to the instance number in training set [51], we split the categories into three disjoint sets: $head$ (more than $10k$), $body$ ($0.5k \sim 10k$) and $tail$ (less than $0.5k$), as demonstrated in Fig.4.

Evaluation Metrics: Due to the reporting bias caused by the data imbalance [41], in this paper, we choose mean Recall $mR@K$ as the evaluation metric instead of the traditional Recall $R@K$. Unlike $R@K$ which only concentrates on common predicate categories and underestimates the informative predicate categories, $mR@K$ averages the recalls across the predicate categories. Following the previous methods, we test the proposed NBP method on three tasks: predicate classification (PredCls), scene graph classification (SGCls) and scene graph detection (SGDet). Specifically, PredCls task aims to predict the predicate labels given the input image, the ground-truth bounding boxes and object labels; SGCls task tries to predict the object and predicate labels given the input image and the ground-truth bounding boxes; SGDet task generates the scene graph from the input image.

Implementation Details: Following [41], [35], in this experiment, ResNeXt-101-FPN [52] (backbone) and Faster-RCNN [53] (object detector) are applied to construct the visual perception module. Like the previous methods, we choose a step training strategy, in which the parameters of the visual

TABLE 1: A performance comparison on Visual Genome dataset.

Method	PredCls		SGCls		SGDet		SGDet(R@100)		
	mR@50	mR@100	mR@50	mR@100	mR@50	mR@100	Head	Body	Tail
RelDN† [48]	15.8	17.2	9.3	9.6	6.0	7.3	34.1	6.6	1.1
Motifs [4]	14.6	15.8	8.0	8.5	5.5	6.8	36.1	7.0	0.0
Motifs* [4]	18.5	20.0	11.1	11.8	8.2	9.7	34.2	8.6	2.1
G-RCNN† [6]	16.4	17.2	9.0	9.5	5.8	6.6	28.6	6.5	0.1
MSDN† [49]	15.9	17.5	9.3	9.7	6.1	7.2	35.1	5.5	0.0
GPS-Net† [8]	15.2	16.6	8.5	9.1	6.7	8.6	34.5	7.0	1.0
GPS-Net†* [8]	19.2	21.4	11.7	12.5	7.4	9.5	30.4	8.5	3.8
VCTree-TDE [41]	25.4	28.7	12.2	14.0	9.3	11.1	24.5	13.9	0.1
BGNN [35]	30.4	32.9	14.3	16.5	10.7	12.6	33.4	13.4	6.4
NBP	**28.5**	**30.6**	**15.1**	**16.5**	**12.9**	**14.7**	**31.7**	**15.0**	**8.9**

• Note: Using bold to represent the proposed NBP method. All the above methods apply ResNeXt-101-FPN as the backbone. $*$ means the re-sampling strategy [33] is applied in this method, and † depicts the reproduced results with the latest code from the authors. In the most difficult yet representative SGDet task, compared with BGNN algorithm, the proposed NBP method improves the $mR@50$ and $mR@100$ performance by $(12.9-10.7)/10.7=20.6\%$ and $(14.7-12.6)/12.6=16.7\%$. Unlike the previous models which mainly detect the common $head$ predicate categories, the proposed NBP method concentrates on detecting the more informative $body$ and $tail$ predicate categories. This could mitigate the biased predicate predication problem caused by the long-tail data distribution. As a result, to further improve the overall SGDet performance, the proposed NBP method achieves much higher detection performance on the more informative $body$ and $tail$ predicate categories at the expense of a relatively low detection performance on the common $head$ predicate categories.

perception module are kept frozen during the training period and we only train the visual context reasoning module. The batch size bs is set to 12. A bi-level data resampling strategy [35] is adopted in this experiment, in which we set the repeat factor $t=0.07$, instance drop rate $\gamma_d=0.7$ and the weight of fusion of the entities features $\rho=-5$. In this paper, we choose different numbers of NBP layers for the above three tasks. Specifically, we employ two NBP layers in the PredCls task and only one NBP layer is applied in the SGCls task. For the SGDet task, we use three NBP layers. An SGD optimizer with the learning rate of $0.008\times bs$ is applied to train the above three tasks.

4.1.2 Comparison with State-of-the-art Methods

Comparison with the original NBP method: For a fair comparison, in this experiment, we compare the proposed NBP method with several state-of-the-art baseline models. Some of them were reproduced using the author's latest codes, while others utilize the original codes but with a specific re-sampling strategy [33]. As demonstrated in Table 1, the proposed NBP method achieves the state-of-the-art performance in the SGCls and SGDet tasks, and comparable performance in PredCls task. Specifically, for the most representative SGDet task, compared with the previous best BGNN model, the proposed NBP method improves the $mR@50$ and $mR@100$ performance by 20.6% and 16.7%.

Moreover, to investigate the biased relationship prediction problem caused by the long-tail data distribution, we compare the $R@100$ performance on the long-tail category groups in the SGDet task in Table 1. For the informative $tail$ and $body$ predicate categories, the proposed NBP method achieves the state-of-the-art performance. In particular, it outperforms the previous methods by a large margin for the most informative $tail$ predicate categories. This implies the proposed NBP method is capable of detecting the informative predicate categories, which are hindered by the problem of having a fewer samples for training. Compared with the previous SGG models, which predominantly detect the common predicate categories, it can achieve relatively unbiased training. To mitigate the biased relationship prediction problem caused by the long-tail data distribution, such unbiased training is much needed for SGG models.

TABLE 2: Performance comparison on the Visual Genome dataset with a balance adjustment strategy.

Method	PredCls		SGCls		SGDet	
	mR@50	mR@100	mR@50	mR@100	mR@50	mR@100
Motifs+BA [40]	29.7	31.7	16.5	17.5	13.5	15.6
VCTree+BA [40]	30.6	32.6	20.1	21.2	13.5	15.7
Transformer+BA [40]	31.9	34.2	18.5	19.4	14.8	17.1
NBP+BA	**35.8**	**37.9**	**20.5**	**21.9**	**14.8**	**17.3**

• Note: All the above methods apply the same balance adjustment strategy as in [40] .

Comparison with the derived NBP+BA method: Following [40], to achieve an even more unbiased training, we adopt a generic balance adjustment strategy in the proposed NBP method, aiming to correct two aspects of imbalance: the semantic space imbalance and the training sample imbalance. For the semantic space level imbalance, a semantic adjustment process is applied to induce the predictions by the NBP method to be more informative by constructing an appropriate transition matrix. For the training sample imbalance, a balanced predicate learning procedure is employed to extend the sampling space for informative predicates. Here, the term informative predicate is used in reference to the Shannon information theory, in which the predicates occurring less frequently are deemed to contain more information. With such simple yet effective information measurement scheme, one can easily generate balanced training samples for the proposed NBP method.

As demonstrated in Fig.5, compared with the NBP method, the resulting NBP+BA algorithm has more bal-

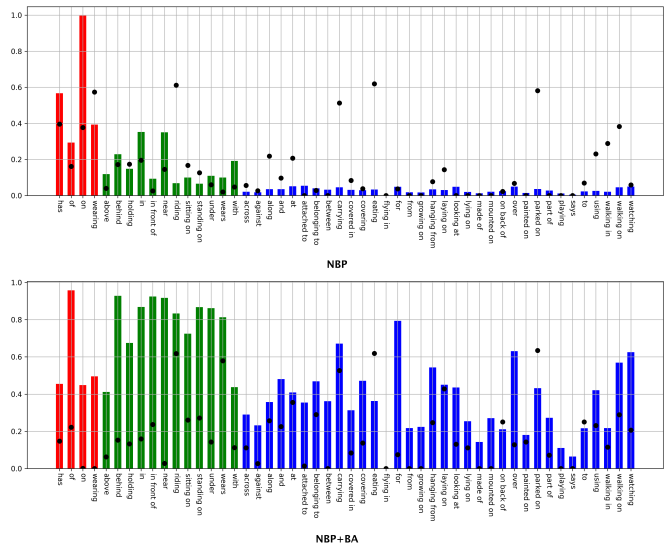


Fig. 5: Comparison of the $mR@100$ performance (represented as black dots) for each predicate category with the proposed NBP method and the resulting NBP+BA algorithm. Here, y axis denotes the min-max normalized frequency. Compared with the NBP method, the resulting NBP+BA algorithm has more balanced training samples, in which the informative $tail$ (blue bars) and $body$ (green bars) predicate categories have comparable training samples as the common $head$ (red bars) predicate categories.

anced training samples, in which the informative $tail$ and $body$ predicate categories have comparable training samples as the common $head$ predicate categories. For the informative $tail$ and $body$ predicate categories, the resulting NBP+BA algorithm generally outperforms the NBP method. In Fig.5, the black dots denote the $mR@100$ performances. It can be seen that the black dots for NBP+BA algorithm are generally higher than the ones for the NBP method.

For a fair comparison, in this experiment, the resulting NBP+BA method is compared with three baseline models as presented in [40]. Specifically, based on the Shannon information theory, the balanced predicate learning procedure discards the redundant training samples of the common $head$ group, and keeps most of the training samples of the informative $body$ and $tail$ groups. As a result, the training samples of the NBP+BA method are more balanced and the resulting data distribution is no longer a long-tail distribution. Moreover, with the transition matrix introduced in the semantic adjustment process, the predictions from the NBP method are further mapped into more informative ones. As demonstrated in Table 2, the resulting NBP+BA method outperforms the previous state-of-the-art methods by a large margin in Visual Genome benchmark, especially for the PredCls task.

4.2 Open Images V6

4.2.1 Experimental Configurations

Benchmark: Open Images V6 [47] is another popular SGG benchmark, which consists of 301 object categories and 31 predicate categories. Compared with Visual Genome, it provides a superior annotation quality. In Open Images V6, there are 126,368 training images, 5322 test images and 1813 validation images. In this experiment, we adopt the same data processing protocols as in [8], [48], [47].

Evaluation Metrics: Following the evaluation protocols in [8], [48], [47], the following evaluation metrics are chosen in this experiment: 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 phrase ($wmAP_{phr}$). Like [8], [47], [48], the weight metric score is defined as: $score_{wtd} = 0.2 \times R@50 + 0.4 \times wmAP_{rel} + 0.4 \times wmAP_{phr}$.

Implementation Details: Following the previous experiment, for the visual perception module, ResNeXt-101-FPN [52] is employed as the backbone and Faster-RCNN [53] is applied as the object detector. As we adopt the step training strategy, freeze the model parameters of the above visual perception module and only train the visual context reasoning module. Moreover, the above bi-level data re-sampling strategy [35] with the same settings is also utilized in this experiment. We set the batch size bs to 12 and use two NBP layers in the visual context reasoning module. An Adam optimizer with learning rate of 0.0001 is applied to train the proposed NBP method.

TABLE 3: A performance comparison on the Open Images V6 dataset.

Method	mR@50	R@50	wmAP_rel	wmAP_phr	score_wtd
RelDN† [48]	33.98	73.08	32.16	33.39	40.84
RelDN†* [48]	37.20	75.34	33.21	34.31	41.97
VCTree† [7]	33.91	74.08	34.16	33.11	40.21
G-RCNN† [6]	34.04	74.51	33.15	34.21	41.84
Motifs† [4]	32.68	71.63	29.91	31.59	38.93
VCTree-TDE† [41]	35.47	69.30	30.74	32.80	39.27
GPS-Net† [8]	35.26	74.81	32.85	33.98	41.69
GPS-Net†* [8]	38.93	74.74	32.77	33.87	41.60
BGNN [35]	40.45	74.98	33.51	34.15	42.06
NBP	**41.97**	**75.54**	**34.44**	**35.66**	**43.08**

* Note: All the above methods apply ResNeXt-101-FPN as the backbone. $*$ means the re-sampling strategy [33] is applied in this method, and † depicts the reproduced results with the latest code from the authors.

4.2.2 Comparison with State-of-the-art Methods

As demonstrated in Table 3, we compare the proposed NBP method with various state-of-the-art methods on the Open Images V6 benchmark. For a fair comparison, in this experiment, most of the baseline models are re-implemented with the author's latest codes. Some of them are used with an additional re-sampling strategy [33]. The proposed NBP method achieves the state-of-the-art performance on all evaluation metrics. Specifically, for the representative $mR@50$ metric, the proposed NBP method outperforms the previous methods by a large margin. Clearly, the above superior performance on the complex Open Images V6 benchmark verifies the effectiveness of the proposed NBP method.

4.3 Ablation Study

In this section, for the proposed NBP method, we first investigate the impact of different types of aggregators on the final scene graph detection performance. Specifically, we compare the SGDet performances of two models: NBP

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3243306

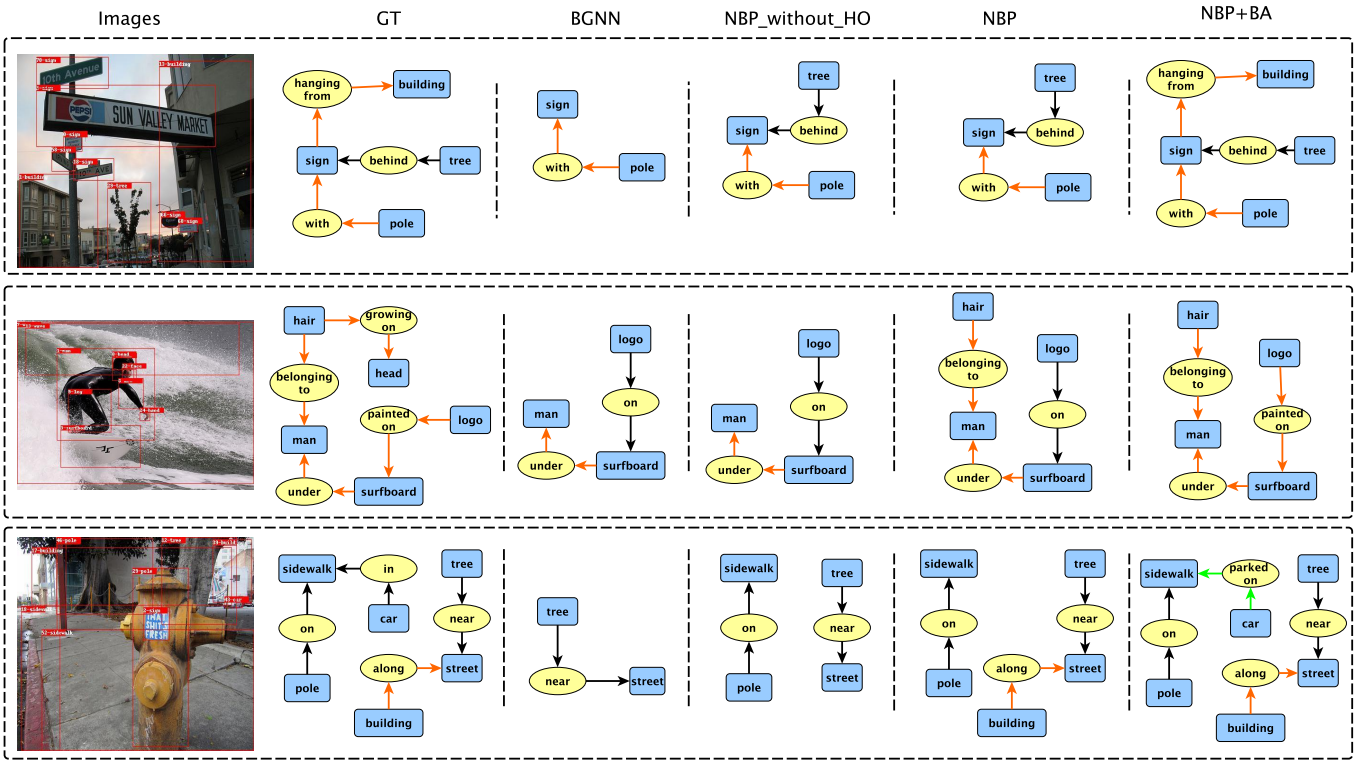


Fig. 6: Visualization of the qualitative results of the ground-truth (GT), the baseline model BGNN, the ablated variant NBP without higher-order dependencies, the proposed NBP method and the derived NBP+BA algorithm in the SGDet task. The black, orange and green arrows represent the triplets with *head* predicate categories, the triplets with *body* or *tail* predicate categories and the reasonable triplets detected by models which are not included in GT, respectively. Compared with the baseline model BGNN, the scene graphs generated by the proposed NBP method and the derived NBP+BA algorithm are much closer to the ground-truth scene graph GT. Moreover, by adding higher-order dependencies, the proposed NBP method could potentially detect more triplet structures.

TABLE 4: An ablation study of different types of aggregators.

Method	Aggregator Type	mR@20	mR@50	mR@100
NBP	max	9.2	12.1	14.2
NBP	**mean**	**10.0**	**12.9**	**14.7**

• Note: Two types of aggregators - max and $mean$ - are compared in this table.

TABLE 5: An ablation study of different types of scoring functions.

Method	Scoring Function	mR@20	mR@50	mR@100
NBP	without HO	7.5	10.5	12.4
NBP	**with HO**	**10.0**	**12.9**	**14.7**

• Note: HO stands for higher order.

TABLE 6: An ablation study of the bias-variance trade-off.

Method	Scoring Function	tS@20	tS@50	tS@100
NBP	without HO	0.105	0.142	0.168
NBP	**with HO**	**0.148**	**0.197**	**0.223**

• Note: HO stands for higher order.

method with the max aggregator and NBP method with $mean$ aggregator, as shown in Table 4. We observe that the NBP method with the $mean$ aggregator constantly outperforms its counterpart NBP algorithm with the max aggregator. This implies that, compared with the max-product method, the sum-product algorithm is more suitable for the scene graph generation task. As a result, in the proposed NBP architecture, we advocate the use of the $mean$ aggregator, instead of the max aggregator, as demonstrated in Equation (15).

Furthermore, we conduct another ablation study to investigate the impact of the proposed scoring function on the final scene graph generation performance. In this study, the baseline model is set to an NBP method with a scoring function containing only unary and pairwise dependencies. As demonstrated in Table 5, the above baseline model is compared with the NBP method employing the proposed scoring function specified in Equation (9). It can be seen that the proposed scoring function with the higher order

dependencies generally produces better performance across all the evaluation metrics, which implies the global contextual information injected by the higher order dependencies play an important role in generating a more consistent interpretation for an input image. This is mainly because the proposed scoring function reduces the model bias, which leads to a better bias-variance trade-off.

In supervised learning, the bias-variance trade-off is a central problem. Ideally, one aims to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. However, in reality, it is generally impossible to achieve both at the same time. High-bias learning algorithms typically produce simpler models that may fail to capture important regularities (i.e. underfit) in the data. In contrast, methods with high variance may be able to represent their training set well but are at risk of overfitting to noisy or unrepresentative training data. As a result, to avoid the above underfitting or overfitting scenarios, one needs to find a better bias-variance trade-off. Specifically, based on the bias–variance decomposition of mean squared error [20], the expected total error is the summation of $bias^2$, $variance$ and $irreducible\ error$, where $bias$ represents the difference between the average prediction of our model and the correct value which we are trying to predict, $variance$ depicts the variability of model prediction for a given data point or a value which tells us spread of our data, $irreducible\ error$ is the error that can not be reduced by creating good models, which is a measure of the amount of noise in our data.

For simplicity, we omit the $irreducible\ error$ in this ablation study and set the total error to be the summation of $bias^2$ and $variance$. In particular, the total error tE is computed as follows:

$$tE = (\mathbb{E}[\hat{f}(x)] - f(x))^2 + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] \quad (16)$$

where $\hat{f}(x)$ represents the predicted scene graph for the input image x, while $f(x)$ denotes the ground-truth scene graph of x. In this paper, we employ a mean recall rate $mR@K$ (where K represents the number of top triplet predictions) to measure the closeness of the prediction to the ground-truth result. Note, $mR@K$ increases when the prediction $\hat{f}(x)$ is closer to the ground-truth result $f(x)$. Therefore, one could reformulate the above equation as:

$$tS@K = (\mathbb{E}[mR@K])^2 + \mathbb{E}[(mR@K - \mathbb{E}[mR@K])^2] \quad (17)$$

where $tS@K$ represents the total similarity when choosing the top K triplet predictions. Specifically, $tS@K$ increases when a model achieves a better bias-variance trade-off. As shown in Table 6, the total similarity $tS@K$ substantially increases if we incorporate the higher-order dependencies into the scoring function, which indicates our proposed NBP model achieves a better bias-variance trade-off.

4.4 Visualization of the Results

To intuitively demonstrate the superiority of the proposed method, in Fig.6, we visually compare the qualitative results of the ground-truth (GT), the baseline model BGNN, the ablated variant NBP without higher-order dependencies, the proposed NBP method and the derived NBP+BA algorithm in the SGDet task. Compared with the baseline model BGNN, the proposed NBP method is capable of detecting more informative $body$ and $tail$ predicates. For instance, the proposed NBP could detect an additional $belonging\ to$ predicate for the middle image. Besides, the spatial informative predicates such as $< under >$ or $< near >$ can also be detected. Moreover, the derived NBP+BA method further improves the above capability. For example, it could detect an additional $hanging\ from$ predicate for the top image, or even a new reasonable predicate $parked\ on$ (which is not included in the ground-truth scene graph GT) for the bottom image. In a word, compared with the baseline model BGNN, the scene graphs generated by the proposed NBP method and the derived NBP+BA algorithm are much closer to the ground-truth scene graph GT. Finally, compared with the ablated variant NBP without the higher-order dependencies, the proposed NBP method with higher-order dependencies was able to detect more meaningful triplets, e.g. the $hair\ belonging\ to\ man$ triplet for the middle image, or the $building\ along\ street$ triplet for the bottom image.

5 Conclusion

In this paper, we proposed a novel neural belief propagation method, which aims to solve two main drawbacks of the previous mean field message passing neural network models, namely that: 1) the output variables are considered to be fully independent within the approximation; 2) only pairwise dependencies are incorporated into the associated scoring function. To find a better bias-variance trade-off, a novel scoring function incorporating higher order dependencies is proposed. The proposed NBP method aims to simulate a classical sum-product algorithm to infer the optimum interpretations for an input image. We validated the proposed generic method on two popular scene graph generation benchmarks: Visual Genome and Open Images V6. The extensive experimental results clearly demonstrate its superiority.

Acknowledgments

This work was supported in part by the U.K. Defence Science and Technology Laboratory, and in part by the 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.

References

[1] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang, "Counterfactual critic multi-agent training for scene graph generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4613–4623.

[2] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6163–6171.

[3] Y. Guo, J. Song, L. Gao, and H. T. Shen, "One-shot scene graph generation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3090–3098.

[4] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.

[5] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 335–351.

[6] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–685.

[7] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6619–6628.

[8] X. Lin, C. Ding, J. Zeng, and D. Tao, "Gps-net: Graph property sensing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3746–3753.

[9] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," *Introduction to statistical relational learning*, vol. 2, pp. 93–128, 2006.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[11] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[12] C. W. Fox and S. J. Roberts, "A tutorial on variational bayesian inference," *Artificial intelligence review*, vol. 38, no. 2, pp. 85–95, 2012.

[13] D. Liu, M. Bober, and J. Kittler, "Visual semantic information pursuit: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 4, pp. 1404–1422, 2019.

[14] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 2008–2026, 2018.

[15] D. Tran, D. Blei, and E. M. Airoldi, "Copula variational inference," *Advances in neural information processing systems*, vol. 28, 2015.

[16] Z. Zhang, F. Wu, and W. S. Lee, "Factor graph neural networks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 8577–8587.

[17] J. S. Yedidia, W. T. Freeman, Y. Weiss *et al.*, "Understanding belief propagation and its generalizations," *Exploring artificial intelligence in the new millennium*, vol. 8, pp. 236–239, 2003.

[18] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on information theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[19] T. Parr, D. Markovic, S. J. Kiebel, and K. J. Friston, "Neuronal message passing using mean-field, bethe, and marginal approximations," *Scientific reports*, vol. 9, no. 1, pp. 1–18, 2019.

[20] R. Kohavi, D. H. Wolpert *et al.*, "Bias plus variance decomposition for zero-one loss functions," in *ICML*, vol. 96, 1996, pp. 275–83.

[21] T. Meltzer, A. Globerson, and Y. Weiss, "Convergent message passing algorithms: a unifying view," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009, pp. 393–401.

[22] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[24] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

[25] D. Teney, L. Liu, and A. van Den Hengel, "Graph-structured representations for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1–9.

[26] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[27] J. Shi, H. Zhang, and J. Li, "Explainable and explicit visual reasoning over scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8376–8384.

[28] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE conference on computer vision and Pattern recognition*, 2017, pp. 3076–3086.

[29] S. Woo, D. Kim, D. Cho, and I. S. Kweon, "Linknet: Relational embedding for scene graph," *Advances in Neural Information Processing Systems*, vol. 31, pp. 560–570, 2018.

[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[31] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 467–482.

[32] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 181–196.

[33] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5356–5364.

[34] X. Hu, Y. Jiang, K. Tang, J. Chen, C. Miao, and H. Zhang, "Learning to segment the tail," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 045–14 054.

[35] R. Li, S. Zhang, B. Wan, and X. He, "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 109–11 119.

[36] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 1567–1578.

[37] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[38] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4367–4375.

[39] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9719–9728.

[40] Y. Guo, L. Gao, X. Wang, Y. Hu, X. Xu, X. Lu, H. T. Shen, and J. Song, "From general to specific: Informative scene graph generation via balance adjustment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 383–16 392.

[41] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3716–3725.

[42] J. Kuck, S. Chakraborty, H. Tang, R. Luo, J. Song, A. Sabharwal, and S. Ermon, "Belief propagation neural networks," *arXiv preprint arXiv:2007.00295*, 2020.

[43] V. G. Satorras and M. Welling, "Neural enhanced belief propagation on factor graphs," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 685–693.

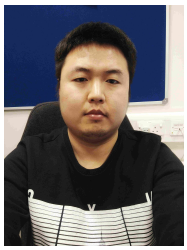
[44] H.-A. Loeliger, J. Dauwels, J. Hu, S. Korl, L. Ping, and F. R. Kschischang, "The factor graph approach to model-based signal processing," *Proceedings of the IEEE*, vol. 95, no. 6, pp. 1295–1322, 2007.

[45] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.

[46] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[47] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *International journal of computer vision*, 2020.

[48] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph parsing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[49] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1261–1270.
[50] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.
[51] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2537–2546.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[53] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.



Daqi Liu is a Research Fellow at 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 (S'94-A'95-M'04) received the M.Sc. and Ph.D. 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 (M'74-LM'12) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished Professor of Machine Intelligence at 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 over 700 scientific papers. His publications have been cited more than 60,000 times (Google Scholar).

He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of Pattern Recognition Letters, Pattern Recognition and Artificial Intelligence, Pattern Analysis and Applications. 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.