In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15816-15826

Visual Distant Supervision for Scene Graph Generation

Yuan Yao[1*], Ao Zhang[1*], Xu Han[1], Mengdi Li[2],
Cornelius Weber[2], Zhiyuan Liu[1†], Stefan Wermter[2], Maosong Sun[1]
[1]Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
[2]Knowledge Technology Group, Department of Informatics, University of Hamburg, Hamburg, Germany
yuan-yao18@mails.tsinghua.edu.cn, zhanga6@outlook.com

Abstract

Scene graph generation aims to identify objects and their relations in images, providing structured image representations that can facilitate numerous applications in computer vision. However, scene graph models usually require supervised learning on large quantities of labeled data with intensive human annotation. In this work, we propose visual distant supervision, a novel paradigm of visual relation learning, which can train scene graph models without any human-labeled data. The intuition is that by aligning commonsense knowledge bases and images, we can automatically create large-scale labeled data to provide distant supervision for visual relation learning. To alleviate the noise in distantly labeled data, we further propose a framework that iteratively estimates the probabilistic relation labels and eliminates the noisy ones. Comprehensive experimental results show that our distantly supervised model outperforms strong weakly supervised and semi-supervised baselines. By further incorporating human-labeled data in a semi-supervised fashion, our model outperforms state-of-the-art fully supervised models by a large margin (e.g., 8.3 *micro- and* 7.8 *macro-recall@50 improvements for predicate classification in Visual Genome evaluation). We make the data and code for this paper publicly available at* `https://github.com/thunlp/VisualDS`.

1. Introduction

Scene graph generation aims to identify objects and their relations in real-world images. For example, the scene graph shown in Figure 1 depicts the image with several relational triples, such as (*person*, `riding`, *horse*) and (*horse*, `standing on`, *beach*). Such structured representations provide deep understanding of the semantic content of images, and have facilitated state-of-the-art models in numerous applications in computer vision, such as visual question answering [17, 42], image retrieval [21, 38], image captioning [50, 12] and image generation [20].

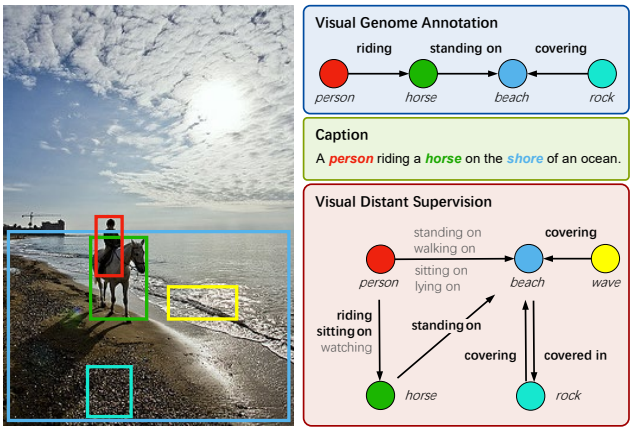

Figure 1. An example from Visual Genome [22] based on the refined relation schemes from Chen *et al*. [5], where human annotation from Visual Genome, weak supervision information from the corresponding caption, and raw relation labels from distant supervision are shown respectively. Correct relation labels are highlighted in bold. By aligning commonsense knowledge bases and images, visual distant supervision can create large-scale labeled data without any human efforts to facilitate visual relation learning. Best viewed in color.

Tremendous efforts have been devoted to generating scene graphs from images [48, 26, 49, 29, 57]. However, scene graph models usually require supervised learning on large quantities of human-labeled data. Manually constructing large-scale datasets for visual relation learning is extremely labor-intensive and time-consuming [29, 22]. Moreover, even with the human-labeled data, scene graph models usually suffer from the long-tail relation distribu-

* indicates equal contribution
† Corresponding author: Z.Liu (liuzy@tsinghua.edu.cn)

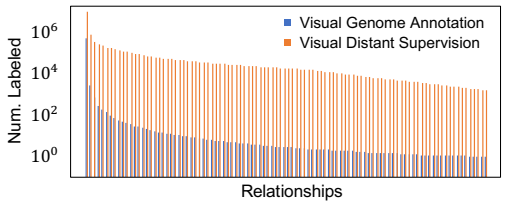

Figure 2. Number of labeled instances of top 3,000 relationships from Visual Genome annotation and visual distant supervision.

tion in real-world scenarios. Figure 2 shows the statistics on Visual Genome [22], where over 98% of the top $3,000$ relation categories do not have sufficient labeled instances and are thus ignored by most scene graph models.

To address the problems, a promising direction is to utilize large-scale unlabeled data with minimal human efforts via semi-supervised or weakly supervised learning. Chen *et al*. [5] propose to first learn a simple relation predictor using several human-labeled seed instances for each relation, and then assign soft labels to unlabeled data to train scene graph models. However, semi-supervised models still require human annotation that scales linearly with the number of relations. Moreover, learning from limited seed instances is vulnerable to high variance and subjective annotation bias. Some works have also explored learning from weakly supervised relation labels, which are obtained by parsing the captions of the corresponding images [58, 32]. However, due to reporting bias [11], captions only summarize a few salient relations in images, and ignore less salient and background relations, e.g., (*rock*, `covering`, *beach*) in Figure 1. The resultant models will thus be biased towards a few salient relations, which cannot well serve scene graph generation that aims to exhaustively extract all reasonable relational triples in the scene.

In this work, we propose *visual distant supervision*, a novel paradigm of visual relation learning, which can train scene graph models without any human-labeled data. The intuition is that commonsense knowledge bases encode relation candidates between objects, which are likely to be expressed in images. For example, as shown in Figure 1, multiple relation candidates, e.g., `riding`, `sitting on` and `watching`, can be retrieved from commonsense knowledge bases for the object pair *person* and *horse*, where `riding` and `sitting on` are actually expressed in the given image. By aligning commonsense knowledge bases and images, we can create large-scale labeled data to provide distant supervision for visual relation learning without any human efforts. Since the distant supervision is provided by knowledge bases, the relations can be exhaustively labeled between all object pairs. We note that many reasonable relation labels from distant supervision are missing in Visual Genome even after intensive human annotations, e.g., (*wave*, `covering`, *beach*) in Figure 1.

Moreover, distant supervision can also alleviate the long-tail problem. As shown in Figure 2, using the same number of images, distant supervision can produce 1-3 orders of magnitude more labeled relation instances than its human-labeled counterpart. Note that the number of distantly labeled relation instances can be arbitrarily large, given the nearly unlimited image data on the Web.

Distant supervision is convenient in training scene graph models without human-labeled data. When human-annotated data is available, distantly labeled data can also be incorporated in a semi-supervised fashion to surpass fully supervised models. We show that after pre-training on distantly labeled data, simple fine-tuning on human-labeled data can lead to significant improvements over strong fully supervised models.

Despite its potential, we note distant supervision may introduce noise in relation labels, e.g., (*person*, `watching`, *horse*) in Figure 1. The reason is that distant supervision only provides relation candidates based on object categories, whereas the actual relations between two objects in an image usually depend on the image content. In principle, the noise in distantly labeled data can be alleviated by maximizing the coherence between distant labels and visual patterns of object pairs. Previous works have shown that without specially designed denoising methods, neural models are capable of detecting noisy labels to some extent, and learning meaningful signals from noisy data [19, 8]. In this work, to better alleviate the noise in distantly labeled data, we further propose a framework that iteratively estimates the probabilistic relation labels and eliminates noisy ones. The framework can be realized by optimizing the coherence of internal statistics of distantly labeled data, and can also be seamlessly integrated with external semantic signals (e.g., image-caption retrieval models), or human-labeled data to achieve better denoising results.

Comprehensive experimental results show that, without using any human-labeled data, our distantly supervised model outperforms strong weakly supervised and semi-supervised baseline methods. By further incorporating human-labeled data in a semi-supervised fashion, our model outperforms state-of-the-art fully supervised models by a large margin (e.g., 8.3 micro- and 7.8 macro-recall@50 improvements for predicate classification task in Visual Genome evaluation). Based on the experiments, we discuss multiple promising directions for future research.

Our contributions are threefold: (1) We propose visual distant supervision, a novel paradigm of visual relation learning, which can train scene graph models without any human-labeled data, and also improve fully supervised models. (2) We propose a denoising framework to alleviate the noise in distantly labeled data. (3) We conduct comprehensive experiments which demonstrate the effectiveness of visual distant supervision and the denoising framework.

2. Related Work

Visual Relation Detection. Identifying visual relations between objects is critical for image understanding, which have received broad attention from the community [29, 57, 13, 14, 7, 35, 3, 52]. Johnson *et al*. [21] further formulate scene graphs that encode all objects and their relations in images into structured graph representations. Tremendous efforts have been devoted to generating scene graphs, including refining contextualized graph features [6, 48, 26, 54], developing computationally efficient scene graph models [25, 49, 58] and designing effective loss functions [49, 59]. However, scene graph models usually require supervised learning on large amounts of human-labeled data [29, 22].

Weakly Supervised Scene Graph Generation. To alleviate the heavy reliance on human-labeled data, recent works on scene graph generation have explored semi-supervised and weakly supervised learning methods. Chen *et al*. [5] propose to bootstrap scene graph models from several human-labeled seed instances for each relation, which still requires manual labor and is vulnerable to high variance. Other works attempt to obtain weakly supervised relation labels from the corresponding image captions. Peyre *et al*. [32] propose to ground the labels to object pairs by imposing global grounding constraints [9]. To improve the computational efficiency, Zhang *et al*. [58] design a network branch to select a pair of object proposals for each relation label. Baldassarre *et al*. [1] propose to first detect the relation via graph networks, and then recover the subject and object of the predicted relation. Zareian *et al*. [53] reformulate scene graphs as bipartite graphs of objects and relations, and align the predicted graphs to their weakly supervised labels. However, since the weakly supervised relation labels are parsed from the corresponding captions, the resultant models will be biased towards the most salient relations, ignoring many less salient and background relations.

Textual Distant Supervision. In natural language processing, there has been a long history of extracting relational triples from text (i.e., textual relation extraction) to complete knowledge bases [18, 31, 56, 60]. Supervised textual relation extraction models are usually limited by the sizes of human-annotated datasets. To address the issue, Mintz *et al*. [30] propose to align Freebase [2], a world knowledge base, to text to provide distant supervision for textual relation extraction. Although both targeting at extracting relations, we provide distant supervision for visual relation learning by aligning commonsense knowledge bases with visual concepts, in contrast to textual distant supervision that aligns world knowledge bases with textual entities.

Learning with Noisy Labels. Visual distant supervision may introduce noisy relation labels, which may hurt the performance of scene graph models. In textual distant supervision, many denoising methods have been developed under the multi-instance learning formulation [55, 28, 61, 15]. However, visual relation detection aims to extract relations on instance level (i.e., predicting relation instances in specific images), whereas textual relation extraction focuses on extracting global relations between entities (i.e., synthesizing information from all sentences containing the entity pair to identify their global relation). Therefore, denoising methods for distantly supervised textual relation extraction cannot well serve visual relation detection. Previous instance-level denoising methods have explored handling noisy labels in image classification [19, 36, 44, 46, 24] and object detection [23, 41, 10] based on internal data statistics. In comparison, our denoising framework cannot only leverage internal data statistics, but can also be seamlessly integrated with external semantic signals and human-labeled data for better denoising results.

3. Problem Definition

We first provide a formal definition of the problem and key terminologies in our work.

Scene Graphs. Formally, a scene graph consists of the following elements: (1) Objects. Each object $obj = (b, c)$ is associated with a bounding box $b \in \mathbb{R}^4$ and a category $c \in \mathcal{C}$, where $\mathcal{C}$ is the object category set. (2) Relations, with $r \in \mathcal{R}$, where $\mathcal{R}$ is the relation category set (including `NA` indicating no relation). Given an image, scene graph models aim to extract the relational triple (s, r, o).

Knowledge Bases. Most knowledge bases store relations between concepts in the form of relational triples (c_i, r, c_j).

Distantly Supervised and Semi-supervised Learning. In the traditional fully supervised relation learning, human labeled data D_L is required to train scene graph models. In distantly supervised relation learning, where no human-labeled data is available, we automatically create distantly labeled data D_S using images and knowledge bases to train scene graph models. When human-labeled data D_L is available, we can further leverage $D_S \cup D_L$ in a semi-supervised fashion to surpass fully supervised models trained on D_L.

4. Visual Distant Supervision

In this section, we introduce the assumption and approach of visual distant supervision, which aims to create large-scale labeled data for visual relation learning.

The key insight of visual distant supervision is that visual relational triples correspond to commonsense knowledge. For example, the relational triple (*person*, `riding`, *horse*) expresses the commonsense "*person can ride horses*". Therefore, commonsense knowledge bases can provide possible relation candidates between visual objects to distantly supervise visual relation learning. To this end, we perform

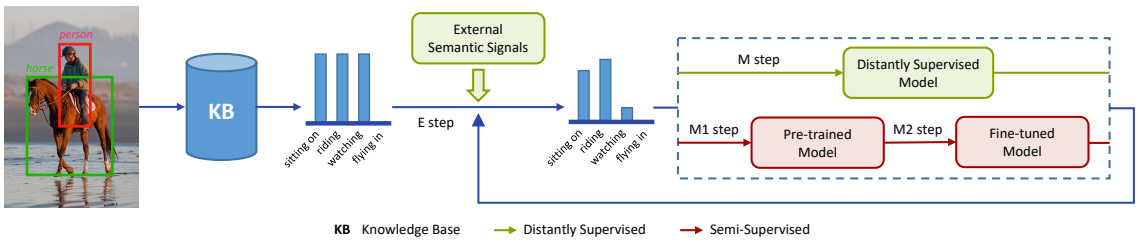

Figure 3. The denoising framework for visual distant supervision. The framework iteratively estimates the probabilistic relation labels based on EM optimization, and can be realized in distantly supervised and semi-supervised fashion. Best viewed in color.

visual distant supervision by first constructing a commonsense knowledge base, and then aligning it to images.

Knowledge Base Construction. Although several commonsense knowledge bases have been constructed [43], we find that they cannot well serve visual distant supervision due to their incompleteness. Therefore, instead of adopting existing knowledge bases, we automatically construct a commonsense knowledge base by extracting relational triples from Web-scale image captions. Specifically, we extract relational triples from Conceptual Captions [40], which contains 3.3M captions of images, using a rule-based textual parser [39]. The resultant knowledge base contains $18,618$ object categories, $63,232$ relation categories and $1,876,659$ distinct relational triples, where each object pair has 1.94 relations on average.

Knowledge Base and Image Alignment. To align the knowledge base and images, we need to obtain the bounding boxes and categories of objects in each image. In this work, we utilize the images and object annotations from Visual Genome, while obtaining object information in open-domain images using object detectors [37] is also applicable. After that, for each object pair, we retrieve all the relation labels in the knowledge base as relation candidates.

Nevertheless, we observe that directly performing distant supervision will produce considerable noise. For example, if there are multiple *person* and *horse* objects in an image, there will be a `riding` relation label between each *person* and *horse* pair. Inspired by previous works [54], we adopt a simple but effective heuristic constraint to filter out a large number of noisy labels. Specifically, we assign distant relation labels for an object pair only if the bounding boxes of the subject and object have overlapping areas. After alignment, the relation labels from distant supervision can cover 70.3% relation labels from Visual Genome.

5. Denoising for Visual Distant Supervision

The relation labels from distant supervision can be readily used to train any scene graph models. However, distant supervision may introduce noisy relation labels, which may hurt the model performance. To alleviate the noise in visual distant supervision, we propose a denoising framework, as shown in Figure 3. Regarding the ground-truth relation labels of distantly labeled data as latent variables, we iteratively estimate the probabilistic relation labels, and eliminate the noisy ones to train any scene graph models. The framework can be realized by optimizing the coherence of internal statistics of distantly labeled data, and can also be seamlessly integrated with external semantic signals (e.g., image-caption retrieval models), or human-labeled data to achieve better denoising results. In this section, we introduce the framework in distantly supervised and semi-supervised settings respectively. We refer readers to the appendix for the pseudo-code of the framework.

5.1. Distantly Supervised Framework

In distantly supervised framework, where only distantly labeled data D_S is available, we aim to refine the probabilistic relation labels of D_S iteratively by maximizing the coherence of its internal statistics using EM optimization.

E step. In the E step of t-th iteration, we estimate the labels of distantly labeled data to obtain $D_S^t = \{(s, \mathbf{r}^t, o)^{(k)}\}_{k=1}^N$, where $\mathbf{r}^t$ indicates the latent relation between the object pair (s, o) in an image.[1] Specifically, $\mathbf{r}^t \in \mathbb{R}^{|\mathcal{R}|}$ is a probabilistic distribution over all relations in $\mathcal{R}$, which comes from either (1) raw labels from distant supervision (in the initial iteration), or (2) probabilistic relation labels estimated by the model. Given an object pair (s, o), we denote the set of retrieved distant labels as $\mathcal{R}_{(s,o)}$. Note that during the EM optimization, we only refine the distant labels in $\mathcal{R}_{(s,o)}$, and keep $\mathbf{r}_i^t = 0$, if $r_i \notin \mathcal{R}_{(s,o)}$.

(1) In the **initial iteration** (i.e., $t = 1$), the relation labels are assigned by aligning knowledge bases and images (see Section 4), denoted as follows:

$$\mathbf{d} = \Psi(s, o, \Lambda), \tag{1}$$

where Λ is the knowledge base, $\Psi(\cdot)$ is the alignment operation. $\mathbf{d}$ is a multi-hot vector where $\mathbf{d}_i = 1$ if $r_i \in \mathcal{R}_{(s,o)}$ and otherwise 0.

We argue that when available, external semantic signals are useful in distinguishing reasonable distant labels from

[1] We omit the superscript k in the following for simplicity.

noisy ones. Without losing generality, in this work, we adopt CLIP [34], a state-of-the-art cross-modal representation model pre-trained on large-scale image-caption pairs[2], to measure the semantic relatedness between a textual relational triple from distant supervision, and the corresponding visual object pair. Specifically, given an object pair, we obtain the visual input by masking the area in the image that is not covered by the bounding boxes of the object pair. To obtain the textual input, we simply concatenate the subject, relation and object in the relational triple into a text snippet. Then the visual and textual inputs are fed into CLIP to obtain their unnormalized relatedness score (i.e., cosine similarity), summarized as follows:

$$\alpha_i = \Phi(v, u_i), \tag{2}$$

where v is the visual input of the object pair, u_i is the textual input of the distantly labeled relation r_i, $\Phi(\cdot,\cdot)$ denotes external semantic signals, and α_i is the relatedness score. After that, we normalize the relatedness score to obtain the probabilistic relation distribution over $\mathcal{R}_{(s,o)}$:

$$\mathbf{e}_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{|\mathcal{R}|} \mathbb{1}[\mathbf{d}_j = 1]\exp(\alpha_j)}, \;\; r_i \in \mathcal{R}_{(s,o)}, \tag{3}$$

where $\mathbb{1}[x]$ is 1 if x is true otherwise 0. $\mathbf{e}$ is the probabilistic relation distribution given by external semantic signals, where $\mathbf{e}_i = 0$ if $r_i \notin \mathcal{R}_{(s,o)}$.

The relation distribution can then be initialized by $\mathbf{r}^1 = \mathbf{e}$. Note that external signals are not necessarily required by the framework (i.e., initialize $\mathbf{r}^1 = \mathbf{d}$ when such external signals are not available).

(2) In the **non-initial iterations** (i.e., $t > 1$), we infer the probabilistic relation distribution by the convex combination of the internal prediction from scene graph models, and external semantic signals:

$$\mathbf{r}_i^t = \omega f_i(s, o; \theta^{t-1}) + (1-\omega)\mathbf{e}_i, \tag{4}$$

where $f_i(s, o; \theta^{t-1})$ is the probability of r_i from the scene graph model with parameter θ^{t-1}. Here $f_i(s, o; \theta^{t-1})$ is obtained by normalizing the relation logits over $\mathcal{R}_{(s,o)}$ as in Equation 3. $\omega \in [0, 1]$ is a weighting hyperparameter, where $\omega = 1$ when external signals are not available.

We note it is possible that none of the distant labels between (s, o) are correct (see (*person*, *beach*) in Figure 1 for example). Therefore, we eliminate noisy object pairs from D_S^t, by discarding object pairs with top $k\%$ NA relation logits given by the scene graph model.

[2]In the distantly supervised framework, we are careful to not introduce human annotated relation data in any component. The knowledge base is constructed using a rule-based method from image captions, and CLIP is pre-trained on image-caption pairs only.

M step. In the M step, given the distant labels from the E step, we optimize the scene graph model parameters θ^{t-1} by maximizing the log-likelihood of D_S^t as follows:

$$\theta^t = \arg\max_{\theta} \mathcal{L}_p(D_S^t; \theta^{t-1}), \tag{5}$$

where $\mathcal{L}_p(D_S^t; \theta^{t-1})$ is the entropy-based log-likelihood function, which incorporates the probabilistic relation distribution of D_S^t in a noise-aware approach as follows:

$$\mathcal{L}_p(D_S^t; \theta^{t-1}) = \sum_{(s,\mathbf{r}^t,o)\in D_S^t} \sum_{i=1}^{|\mathcal{R}|} \mathbf{r}_i^t (\mathbb{1}[\mathbf{d}_i = 1]\log f_i(s, o; \theta^{t-1}) + \mathbb{1}[\mathbf{d}_i = 0]\log(1 - f_i(s, o; \theta^{t-1}))), \tag{6}$$

where θ^0 is randomly initialized.

5.2. Semi-supervised Framework

Distantly supervised models can be further integrated with human-labeled data to surpass fully supervised models. In fact, we find after pre-training on distantly supervised data (see Section 5.1), simple fine-tuning on human-labeled data can lead to significant improvements over fully supervised models. This simple pre-training and fine-tuning paradigm is appealing, since it does not change the *number of parameters*, *architectures* and *overhead* in training specific downstream scene graph models.

Nevertheless, we find closely integrating human-labeled data in the denoising framework can yield better performance, since coherence can be achieved between distantly labeled data D_S and human-labeled data D_L for mutual enhancement. Our semi-supervised framework largely follows the distantly supervised framework in Section 5.1, where we estimate probabilistic relation labels in E step, and optimizing model parameters in M step. To integrate human-labeled data, we further decompose the M step into two sub-steps: pre-training on distantly labeled data (M1 step), and fine-tuning on human-labeled data (M2 step).

E step. In the E step of t-th iteration, we estimate the labels of distantly supervised data D_S^t. Here $\mathbf{r}_t$ is obtained by (1) first obtaining raw distant labels $\mathbf{d}$ as in Equation 1, and (2) then estimating probabilistic relation labels by the fine-tuned scene graph model as follows:

$$\mathbf{r}_i^t = f_i(s, o; \theta_2^{t-1}), \;\; r_i \in \mathcal{R}_{(s,o)}, \tag{7}$$

where $f_i(\cdot; \theta_2^{t-1})$ is the fine-tuned scene graph model from the M2 step of the previous iteration. In the initial iteration, $f_i(\cdot; \theta_2^0)$ is initialized by a fully supervised model. Note Equation 7 does not include external semantic signals, since models fine-tuned on human-labeled data can provide more direct denoising signals. After that, we discard noisy object pairs (see Section 5.1). To better cope with the fine-tuning

	Models	Predicate Classification				Scene Graph Classification				Scene Graph Detection				Mean
		R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	
Baselines	Freq [54]*	20.80	20.98	-	-	10.92	11.08	-	-	11.01	11.64	-	-	-
	Freq-Overlap [54]*	20.90	22.21	-	-	9.91	9.91	-	-	10.84	10.86	-	-	-
	Decision Tree [33]*	33.02	33.35	-	-	14.51	14.57	-	-	12.58	13.23	-	-	-
	Label Propagation [62]*	25.17	25.41	-	-	9.91	9.97	-	-	6.74	6.83	-	-	-
	Weak Supervision†	44.96	47.19	24.58	27.14	19.27	19.93	6.97	7.54	19.78	21.33	5.01	5.41	20.76
	Limited Labels [5]	49.68	50.73	37.43	38.91	24.65	25.08	13.30	13.94	22.87	24.16	12.66	13.39	27.23
DS (Ours)	EXT	6.64	9.74	10.66	15.16	3.96	4.82	4.25	4.92	1.93	3.06	1.66	2.49	5.77
	Raw Label	30.61	33.48	20.98	23.25	15.69	16.99	11.06	12.53	9.36	10.26	6.56	7.13	16.49
	Raw Label + EXT	38.21	40.90	24.94	27.45	17.52	18.85	11.66	12.56	15.84	18.31	9.49	11.23	20.58
	Motif†	48.88	51.73	34.40	39.69	23.15	24.18	15.81	16.66	18.73	22.10	10.89	13.34	26.63
	Motif	50.23	53.18	33.99	40.62	24.90	26.00	16.50	18.03	20.09	22.74	12.21	14.42	27.74
	Motif + DNS + EXT	**53.40**	**56.54**	**37.68**	**41.98**	**26.12**	**27.46**	**17.20**	**18.39**	**23.69**	**25.59**	**13.84**	**15.23**	**29.76**
FS	Motif [54]	67.93	70.20	52.65	55.41	31.14	31.92	23.53	25.27	28.90	31.25	18.26	20.63	38.09
SS	Motif + Pretrain (Ours)	73.22	75.04	**60.44**	**63.67**	34.11	34.88	26.51	27.94	30.70	33.32	**24.76**	27.45	42.67
	Motif + DNS (Ours)	**76.28**	**77.98**	60.20	63.61	**35.93**	**36.47**	**28.07**	**30.09**	**33.94**	**37.26**	23.90	**28.06**	**44.31**

Table 1. Main results of visual distant supervision (%). DS: distantly supervised, FS: fully supervised, SS: semi-supervised. EXT: external semantic signal, DNS: denoising. * denotes results from Chen *et al*. [5], † indicates models trained on the images with captions.

procedure on human-labeled data, where models are usually optimized towards a single discrete relation label between an object pair, we discretize $\mathbf{r}^t$ into a one-hot vector $\hat{\mathbf{r}}^t$, where $\hat{\mathbf{r}}^t_i = 1$, if $i = \arg\max_j \mathbf{r}^t_j$.

M1 step. After obtaining the distant label $\hat{\mathbf{r}}^t$ from the E step, we pre-train the scene graph model from scratch: $\theta^t_1 = \arg\max_\theta \mathcal{L}_q(D^t_S; \theta)$, where $\mathcal{L}_q$ is the cross-entropy based objective as follows:

$$\mathcal{L}_q(D^t_S; \theta) = \sum_{(s,\hat{\mathbf{r}}^t,o)\in D^t_S} \sum_{i=1}^{|\mathcal{R}|} \mathbb{1}[\hat{\mathbf{r}}^t_i = 1] \log f_i(s, o; \theta). \quad (8)$$

M2 step. In the M2 step, we simply fine-tune the pre-trained scene graph model on the human labeled data with $\theta^t_2 = \arg\max_\theta \mathcal{L}_q(D_L; \theta^t_1)$.

6. Experiments

In this section, we empirically evaluate visual distant supervision and the denoising framework on scene graph generation. We also show the advantage of visual distant supervision in dealing with long-tail problems, and its promising potential when equipped with ideal knowledge bases.

6.1. Experimental Settings

We first introduce the experimental settings, including datasets, evaluation metrics and baselines.

Datasets. We evaluate our models on Visual Genome [22], a widely adopted benchmark for scene graph generation [48, 54, 5, 53]. Each image in the dataset is manually annotated with objects (bounding boxes and object categories) and relations. In our experiments, during training distant supervision is performed using the intersection of relations from Visual Genome and the knowledge base. During evaluation, in the main experiments, we adopt the refined relation schemes and data split from Chen *et al*. [5], which removes hypernyms and redundant synonyms in the most frequent 50 relation categories in Visual Genome, resulting in 20 well-defined relation categories. We also report experimental results on the Visual Genome dataset with 50 relation categories in appendix. We refer readers to the appendix for more details about data statistics.

Evaluation Metrics. Following previous works [48, 54, 5], we assess our approach in three standard evaluation modes: (1) Predicate classification. Given the ground-truth bounding boxes and categories of objects in an image, models need to predict the predicates (i.e., relations) between object pairs. (2) Scene graph classification. Given the ground-truth bounding boxes of objects, models need to predict object categories and relations. (3) Scene graph detection. Given an image, models are asked to predict bounding boxes and categories of objects, and relations between objects. We adopt the widely used micro-recall@K (**R@K**) metric to evaluate the model performance [48, 54, 5], which calculates the recall in top K relation predictions. To investigate the model performance in dealing with long-tail relations, we also report macro-recall@K (**mR@K**) [4, 45], which calculates the mean recall of all relations in top K predictions. Following Zellers *et al*. [54], we also report the mean of these metrics to show the overall performance.

Baselines. We compare our models with strong baselines. (1) The first series of baselines learn visual relations from a few (i.e., 10) human-labeled seed instances for each relation. Frequency-based baseline (**Freq**) [54] predicts the most frequent relation between an object pair. Overlap-enhanced frequency-based baseline (**Freq-Overlap**) [54]

	Models	Predicate Classification				Scene Graph Classification				Scene Graph Detection				Mean
		R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	
DS	Motif	50.23	53.18	33.99	40.62	24.90	26.00	16.50	18.03	20.09	22.74	12.21	14.42	27.74
DS	Motif + Cleanness Loss [23]	51.10	54.23	34.69	42.67	24.06	24.98	16.46	18.56	21.94	23.89	13.21	14.49	28.36
DS	Motif + DNS (iter 1)	50.23	53.18	33.99	40.62	24.90	26.00	16.50	18.03	20.09	22.74	12.21	14.42	27.74
DS	+ DNS (iter 2)	51.54	54.53	36.93	41.97	24.81	26.08	16.13	17.56	22.83	24.36	13.48	14.45	28.72
DS	Motif + DNS + EXT (iter 1)	52.82	55.98	36.25	41.66	25.79	26.98	**17.39**	**18.56**	22.63	25.12	13.30	**15.40**	29.32
DS	+ DNS + EXT (iter 2)	**53.40**	**56.54**	**37.68**	**41.98**	**26.12**	**27.46**	17.20	18.39	**23.69**	**25.59**	**13.84**	15.23	**29.76**
FS	Motif [54]	67.93	70.20	52.65	55.41	31.14	31.92	23.53	25.27	28.90	31.25	18.26	20.63	38.09
SS	Motif + DNS (iter 1)	73.50	75.33	**61.40**	**65.20**	35.39	35.98	**28.71**	**30.25**	**34.83**	**37.68**	**24.78**	27.90	44.25
SS	+ DNS (iter 2)	**76.28**	**77.98**	60.20	63.61	**35.93**	**36.47**	28.07	30.09	33.94	37.26	23.90	**28.06**	**44.31**

Table 2. Experimental results of denoising visual distant supervision (%). Results of different denoising iterations are shown. DS: distantly supervised, FS: fully supervised, SS: semi-supervised. EXT: external semantic signal, DNS: denoising.

further filters out non-overlapping object pairs. Following Chen *et al*. [5], we also compare with learning **Decision Tree** [33] from seed instances. (2) For *semi-supervised methods* that further incorporate unlabeled data, following Chen *et al*. [5], we compare with **Label Propagation** [62] that propagates the labels of the seed data to unlabeled data based on data point communities. **Limited Labels** [5] is the state-of-the-art semi-supervised scene graph model, which first learns a relational generative model using seed instances, and then assigns soft labels to unlabeled data to train scene graph models. (3) We also compare with strong *weakly supervised models* (**Weak Supervision**†) that are supervised by the relation labels parsed from the captions of the corresponding images [32, 58]. Specifically, we use images with captions in Visual Genome to train the weakly supervised model. We label the object pairs with relations parsed from the corresponding caption, and employ the overlapping constraint to filter out noisy labels (see Section 4). For fair comparisons, we also train a distantly supervised model without denoising based on the same images with captions in Visual Genome (**Motif**†). (4) For *fully supervised methods*, we compare with the strong and widely adopted Neural Motif (**Motif**) [54]. For fair comparisons, all the neural models in our experiments are implemented based on the Neural Motif model, with ResNeXt-101-FPN [27, 47] as the backbone. (5) For *denoising baselines*, we adapt **Cleanness Loss** [23] that heuristically down-weight the relation labels with large loss. We refer readers to the appendix for more implementation details.

Ablations. To investigate the contribution of each component, we conduct ablation study. (1) In *distantly supervised setting*, we perform distant supervision based on the general knowledge base from Section 4. **Raw Label** predicts relations by raw distant relation labels. **EXT** indicates external semantic signals. **Motif** denotes training on raw distant relation labels, and **DNS** denotes denoising based on the proposed framework. (2) In *semi-supervised setting*, in addition to distantly labeled data, we assume access to full human-annotated relation data. **Pretrain** indicates directly fine-tuning the model pre-trained on distantly supervised data from the general knowledge base. **DNS** denotes denoising based on the targeted knowledge base constructed from Visual Genome training annotations.

6.2. Effect of Visual Distant Supervision

We report the main results of visual distant supervision in Table 1, from which we have the following observations: (1) Without using any human-labeled data, our denoised distantly supervised model significantly outperforms all baseline methods, including weakly supervised methods and even strong semi-supervised approaches that utilize human-labeled seed data. (2) By further incorporating human-labeled data, our semi-supervised models consistently outperform state-of-the-art fully supervised models by a large margin, e.g., 8.3 R@50 improvement for predicate classification. Specifically, simple fine-tuning of the pre-trained model can lead to significant improvement. Since the model is pre-trained on general knowledge bases, it can also be directly fine-tuned on any other scene graph dataset to achieve strong performance. Moreover, by closely integrating human-labeled data and distantly labeled data in the denoising framework, we can achieve even better performance. (3) Notably, our models achieve competitive macro-recall, which shows that our models are not biased towards a few frequent relations, and can better deal with the long-tail problem. In summary, visual distant supervision can effectively create large-scale labeled data to facilitate visual relation learning in both distantly supervised and semi-supervised scenarios.

6.3. Effect of the Denoising Framework

The experimental results of denoising distant supervision are shown in Table 2, from which we observe that: Equipped with the proposed denoising framework, our models show consistent improvements over baseline models in both distantly supervised and semi-supervised settings. Specifically, in distantly supervised setting, the model performance improves with the iterative optimization of the coherence of internal data statistics. Further incorporating external semantic signals and human-labeled data cannot

	Models	R@50	R@100	mR@50	mR@100
DS (Ours)	Raw Label	35.62 (+5.01)	39.78 (+6.30)	34.83 (+13.85)	39.45 (+16.20)
	Raw Label + EXT	45.07 (+6.86)	49.00 (+8.10)	44.18 (+19.24)	48.56 (+21.11)
	Motif	53.02 (+2.79)	56.31 (+3.13)	46.65 (+12.66)	50.90 (+10.28)
	Motif + DNS + EXT	**55.54** (+2.14)	**58.99** (+2.45)	**50.87** (+13.19)	**55.69** (+13.71)
FS	Motif [54]	67.93	70.02	52.65	55.41

Table 3. Experimental results of distant supervision with ideal knowledge bases on predicate classification (%). We also show the absolute improvements with respect to the results from general knowledge bases. DS: distantly supervised, FS: fully supervised.

only boost the denoising performance, but also speed up the convergence of the iterative algorithm. The reason is that external semantic signals and human-labeled data can provide strong auxiliary denoising signals for both better initialization and iteration of the framework. The results show that the proposed denoising framework can effectively alleviate the noise in visual distant supervision in both distantly supervised and semi-supervised settings.

6.4. Analysis

Distant Supervision with Ideal Knowledge Bases. The effectiveness of visual distant supervision may be limited by the incompleteness of knowledge bases, and mismatch in relation and object names between knowledge bases and images. We show the potential of visual distant supervision with ideal knowledge bases. Specifically, we construct an ideal knowledge base from the training annotations of Visual Genome, which better covers the relational knowledge in the dataset, and can be well aligned to Visual Genome images. Experimental results are shown in Table 3, from which we observe that: Equipped with ideal knowledge bases, the performance of distantly supervised models improves significantly. Notably, the macro-recall of the denoised distantly supervised model dramatically improves, e.g., with 13.2 absolute gain in mR@50, *achieving comparable macro performance to fully supervised approaches*. Therefore, we expect visual distant supervision will even better promote visual relation learning, given that knowledge bases are becoming increasingly complete.

Human Evaluation. Since relations in existing scene graph datasets are typically not exhaustively annotated, previous works have concentrated on evaluating models with recall metric [48, 54, 5]. To provide multi-dimensional evaluation, we further evaluate the precision of our scene graph models. Specifically, we randomly sample 200 images from the validation set of Visual Genome. For each image, we obtain 20 top-ranking relational triples extracted by a scene graph model. Then we ask human annotators to label whether each relational triple is correctly identified from the image. We show the human evaluation results in Table 4, where micro- and macro-precision@K are reported. (1) For raw labels from distant supervision, in addition to the 10.9% strictly correct labels in top 20 recommendations, we find that 32.6% of the distant labels are plausible, which also provides useful signals for visual relation learning. (2) The denoised distantly supervised model achieves competitive precision. Moreover, compared to recall evaluation, our semi-supervised model exhibits more significant relative advantage over the fully supervised model. The results show that with appropriate denoising, distant supervision can lead scene graph models to strong performance in both precision and recall.

	Models	P@10	P@20	mP@10	mP@20
DS	Raw Labels	12.07	10.85	11.41	12.04
	Motif + DNS + EXT	**31.93**	**25.29**	**24.79**	**20.79**
FS	Motif [54]	42.22	31.09	39.60	29.22
SS	Motif + DNS	**50.58**	**38.68**	**47.49**	**38.52**

Table 4. Human evaluation results on predicate classification (%), where micro- and macro-precision@K are reported. DS: distantly supervised, FS: fully supervised, SS: semi-supervised.

7. Discussion and Outlook

In this work, we show the promising potential of visual distant supervision in visual relation learning. In the future, given the recent advances in textual relation extraction [28, 16, 51], we believe the following research directions are worth exploring: (1) Developing more sophisticated denoising methods to better realize the potential of visual distant supervision. (2) Completing commonsense knowledge bases by extracting *global* relations between object pairs from multiple images. (3) Reducing the overhead and bias in annotating large-scale visual relation learning datasets based on raw labels from distant supervision.

8. Conclusion

In this work, we propose visual distant supervision and a denoising framework for visual relation learning. Comprehensive experiments demonstrate the effectiveness of visual distant supervision and the denoising framework. In this work, we denoise distant labels for each object pair in isolation. In the future, we will explore modeling the holistic coherence of the produced scene graph to better alleviate the noise in visual distant supervision. It is also important to investigate whether visual distant supervision will introduce extra bias to scene graph models.

9. Acknowledgement

This work is jointly funded by the Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) in Project Crossmodal Learning, NSFC 62061136001 / DFG TRR-169. Yao is also supported by 2020 Tencent Rhino-Bird Elite Training Program.

References

[1] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *Proceedings of ECCV*, pages 612–630, 2020.

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD*, pages 1247–1250, 2008.

[3] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of CVPR*, pages 1017–1025, 2015.

[4] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of CVPR*, pages 6163–6171, 2019.

[5] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of CVPR*, pages 2580–2590, 2019.

[6] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of CVPR*, pages 3076–3086, 2017.

[7] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. In *Proceedings of ECCV*, pages 158–172, 2012.

[8] Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. Reinforcement learning for relation classification from noisy data. In *Proceedings of AAAI*, volume 32, 2018.

[9] James R. Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010.

[10] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. NOTE-RCNN: NOise Tolerant Ensemble RCNN for semi-supervised object detection. In *Proceedings of CVPR*, pages 9508–9517, 2019.

[11] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, pages 25–30, 2013.

[12] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of CVPR*, pages 10323–10332, 2019.

[13] Abhinav Gupta and Larry S Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proceedings of ECCV*, pages 16–29, 2008.

[14] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

[15] Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of EMNLP*, pages 2236–2245, 2018.

[16] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of EMNLP*, pages 4803–4809, 2018.

[17] Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Proceedings of NeurIPS*, pages 5901–5914, 2019.

[18] Scott B Huffman. Learning information extraction patterns from examples. In *International Joint Conference on Artificial Intelligence*, pages 246–260, 1995.

[19] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313. PMLR, 2018.

[20] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of CVPR*, pages 1219–1228, 2018.

[21] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of CVPR*, pages 3668–3678, 2015.

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[23] Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, and Larry S Davis. Learning from noisy anchors for one-stage object detection. In *Proceedings of CVPR*, pages 10588–10597, 2020.

[24] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of CVPR*, pages 5051–5059, 2019.

[25] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of ECCV*, pages 335–351, 2018.

[26] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of CVPR*, pages 1261–1270, 2017.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of CVPR*, pages 2117–2125, 2017.

[28] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, pages 2124–2133, 2016.

[29] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of ECCV*, pages 852–869, 2016.

[30] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, pages 1003–1011, 2009.

[31] R Mooney. Relational learning of pattern-match rules for information extraction. In *Proceedings of NCAI*, volume 328, page 334, 1999.

[32] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of CVPR*, pages 5179–5188, 2017.

[33] J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2.

[35] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Charles Rosenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *Proceedings of CVPR*, pages 1100–1109, 2015.

[36] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of ICML*, pages 4334–4343. PMLR, 2018.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of NeurIPS*, pages 91–99, 2015.

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

[39] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 70–80, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, pages 2556–2565, 2018.

[41] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Xiaopeng Hong, Feng Zheng, Jianzhuang Liu, Mingliang Xu, and Qi Tian. Noise-aware fully webly supervised object detection. In *Proceedings of CVPR*, pages 11326–11335, 2020.

[42] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of CVPR*, pages 8376–8384, 2019.

[43] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, volume 31, 2017.

[44] Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of NeurIPS*, 2018.

[45] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of CVPR*, pages 6619–6628, 2019.

[46] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. Combating label noise in deep learning using abstention. In *Proceedings of ICML*, pages 6234–6243. PMLR, 2019.

[47] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of CVPR*, pages 1492–1500, 2017.

[48] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of CVPR*, pages 5410–5419, 2017.

[49] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *Proceedings of ECCV*, pages 670–685, 2018.

[50] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of CVPR*, pages 10685–10694, 2019.

[51] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of ACL*, pages 764–777, 2019.

[52] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of CVPR*, pages 5534–5542, 2016.

[53] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of CVPR*, pages 3736–3745, 2020.

[54] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of CVPR*, pages 5831–5840, 2018.

[55] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, pages 1753–1762, 2015.

[56] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, pages 2335–2344, 2014.

[57] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of CVPR*, pages 5532–5540, 2017.

[58] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. PPR-FCN: Weakly supervised visual relation detection via parallel pairwise R-FCN. In *Proceedings of CVPR*, pages 4233–4241, 2017.

[59] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of CVPR*, pages 11535–11543, 2019.

[60] Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. Bidirectional long short-term memory networks for relation classification. In *Proceedings of PACLIC*, pages 73–78, 2015.

[61] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of EMNLP*, pages 35–45, 2017.

[62] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

Appendix: Visual Distant Supervision for Scene Graph Generation

1. The Denoising Framework: Pseudo-Code

In this section, we provide the pseudo-code of the denoising framework in distantly supervised and semi-supervised settings respectively.

Algorithm 1 Distantly Supervised Denoising Framework

Require: Λ: commonsense knowledge base
Require: D_S: distantly labeled image data
Require: $f(\cdot;\theta)$: any scene graph model, with parameter θ
Optional: Φ: external semantic signal
Randomly initialize the model parameter θ^0
// Initial E step: estimate the probabilistic distant labels
Obtain distant labels $\mathbf{d} = \Psi(s, o, \Lambda)$
if external signal Φ available **then**
    Initialize $\mathbf{r}^1 = \mathbf{e}$
else
    Initialize $\mathbf{r}^1 = \mathbf{d}$
end if
// Initial M step: optimize model parameter
Optimize $\theta^1 = \arg\max_\theta \mathcal{L}(D_S^1; \theta^0)$
while not done **do**
    // E step: estimate the probabilistic distant labels
    if external signal Φ available **then**
        Estimate $\mathbf{r}_i^t = \omega f_i(s, o; \theta^{t-1}) + (1-\omega)\mathbf{e}_i$
    else
        Estimate $\mathbf{r}_i^t = f_i(s, o; \theta^{t-1})$
    end if
    Eliminate noisy object pairs
    // M step: optimize model parameter
    Optimize $\theta^t = \arg\max_\theta \mathcal{L}_p(D_S^t; \theta^{t-1})$
end while

2. Implementation Details

In this section, we provide implementation details of our model and baseline methods. For fair comparisons, all the neural models in our experiments are implemented using the same object detector, scene graph model and backbone.

Object Detector. We adopt the object detector implementation from Tang *et al*. [6]. Specifically, the object detector is trained using SGD optimizer with learning rate 8×10^{-3} and batch size 8. During the training process, the learning rate is decreased two times by 10 in $30,000$ and $40,000$ iterations respectively.

Algorithm 2 Semi-Supervised Denoising Framework

Require: Λ: commonsense knowledge base
Require: D_S: distantly labeled image data
Require: D_L: human-labeled image data
Require: $f(\cdot;\theta)$: any scene graph model, with parameter θ
Initialize $f(\cdot;\theta^0)$ with fully supervised model $\theta^0 = \arg\max_\theta \mathcal{L}_q(D_L; \theta)$
while not done **do**
    // E step: estimate the probabilistic distant labels
    Estimate $\mathbf{r}_i^t = f_i(s, o; \theta_2^{t-1})$
    Eliminate noisy object pairs
    // M1 step: pre-train on distantly labeled data D_S^t
    Optimize $\theta_1^t = \arg\max_\theta \mathcal{L}_q(D_S^t; \theta)$
    // M2 step: fine-tune on human-labeled data D_L
    Optimize $\theta_2^t = \arg\max_\theta \mathcal{L}_q(D_L; \theta_1^t)$
end while

Scene Graph Model. For the base scene graph model, we follow the implementation of Neural Motif [8], with ResNeXt-101-FPN [4, 7] as the backbone. For predicate classification and scene graph classification, the ratio of positive relation samples and negative relation samples in each image is at most 1:3. For scene graph detection, a relational triplet is considered as positive only if the detected object pairs match ground-truth annotations, i.e., with identical object categories and bounding box IoU > 0.5. We find that this strict constraint leads to sparse positive supervision in experiments, especially in our distantly supervised setting. To address the issue, we change the ratio of positive and negative relation samples in distantly supervised setting to strictly 1:1. During evaluation, we only keep 64 object bounding box predictions. The models are trained using SGD optimizer on 2 NVIDIA GeForce RTX 2080 Ti, with momentum 0.9, batch size 12 and weight decay 5×10^{-4}.

Our Model. All the hyperparameters of our model are selected by grid search on the validation set. (1) In *distantly supervised setting*, for the denoising framework, the weighting hyperparameter ω is 0.9, and 75% noisy object pairs are discarded. In the first iteration, we train the model with learning rate 0.12, and decrease the learning rate 3 times after the plateaus of validation performance. In the second iteration, the learning rate is 0.012, and decays 1 times after the validation performance plateaus. Note that following Devlin *et al*. [2], the learning rate in the second iteration

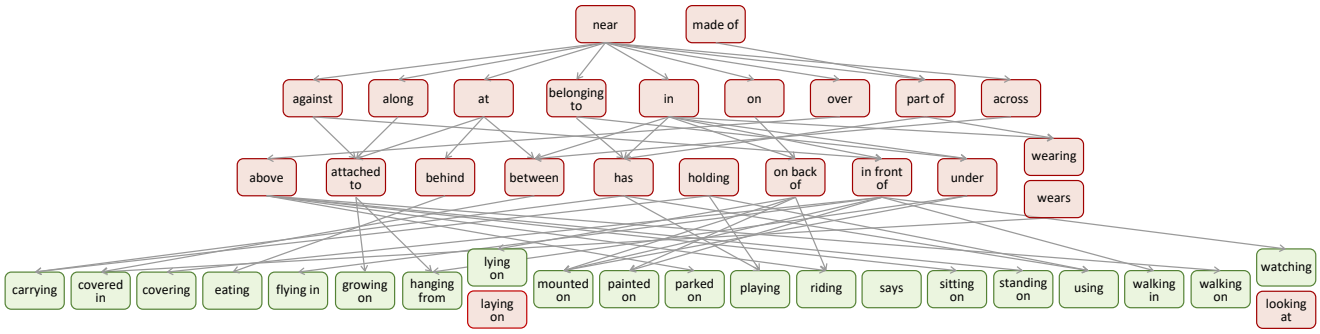

Figure 1. Dependencies of Visual Genome relations from Chen *et al*. [1]. Directed arrows: hypernyms. Stacked nodes: synonyms. Red nodes: removed relations. Green nodes: retained relations.

	Models	Predicate Classification				Scene Graph Classification				Scene Graph Detection				Mean
		R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	R@50	R@100	mR@50	mR@100	
DS (Ours)	Raw Label	16.93	19.02	5.75	7.15	11.62	12.59	4.01	4.64	7.52	7.79	2.32	2.49	8.49
DS (Ours)	Motif	33.21	36.17	**10.84**	**12.48**	20.35	21.85	**5.23**	**5.91**	12.89	15.48	**4.51**	**5.58**	15.38
DS (Ours)	Motif + DNS	35.53	38.28	9.35	10.74	21.33	22.70	4.79	5.36	15.04	17.86	3.83	4.66	15.79
DS (Ours)	Motif + DNS + EXT	**36.43**	**39.21**	8.68	10.03	**21.88**	**23.21**	3.80	4.23	**16.32**	**18.78**	3.82	4.55	**15.91**
FS	Motif [8]	63.96	65.93	15.15	16.24	38.04	38.90	8.66	9.25	**31.00**	35.06	6.66	7.73	28.05
SS	Motif + DNS (Ours)	**64.43**	**66.43**	**16.12**	**17.47**	**38.38**	**39.25**	**9.27**	**9.86**	30.91	**35.08**	**7.03**	**8.29**	**28.54**

Table 1. Results of visual distant supervision on Visual Genome 50 predicates (%). DS: distantly supervised, SS: semi-supervised.

is smaller than the first iteration, since we are actually fine-tuning the model parameter inherited from the first iteration. (2) In *semi-supervised setting*, for the denoising framework, no object pairs are discarded. The initial fully supervised model is trained with learning rate 0.12. In both iterations, the learning rate is 0.12 for pre-training, and 0.012 for fine-tuning. The learning rate decays 2 and 1 times in the first and second iterations respectively. To fine-tune the pre-trained distantly supervised model without semi-supervised denoising, we optimize with learning rate 0.012 that decays 2 times. The decay rate is 10 for all models.

Baselines. For the Limited Labels [1], we train the decision trees for 200 different trails on 10 randomly sampled human-labeled seed instances for each relation, and select the best models according to the performance on the validation set. For the weakly supervised model, since Visual Genome does not have image-level captions, we utilize all the images in Visual Genome training set that have captions from COCO [5], resulting in $35,340$ images with captions in total. Then we train the weakly supervised model with all Visual Genome object annotations from these images, and relation labels parsed from the corresponding captions. For the Cleanness Loss [3], we denoise with soft weight given by the confidence of the scene graph model.

3. Data statistics

In our main experiments, we adopt the refined relation schemes from Chen *et al*. [1], which removes hypernyms (e.g., `near` and `on`), redundant synonyms (e.g., `lying on` and `laying on`), and unclear relations (e.g., `and`) in the most frequent 50 relation categories in Visual Genome, resulting in 20 well-defined relation categories. The relation dependencies from Chen *et al*. [1] are shown in Figure 1. The dataset contains $10,986$, $1,566$ and $3,025$ images in training, validation and test set respectively, where each image contains an average of 13.58 objects, 2.10 human-labeled relation instances and 15.60 distantly labeled relation instances.

4. Supplementary Experiments

Case Study. We provide qualitative examples in Figure 2 for better understanding of different scene graph models.

Results on 50 Visual Genome Predicates. We report the experimental results on 50 Visual Genome predicates in Table 1. We observe that although reasonable performance can be achieved, the improvement from distant supervision and the denoising framework shrinks. This is because that the 50 relations are not well-defined, where the major relations are problematic hypernym (e.g., `near` and `on`), redundant synonym (e.g., `lying on` and `laying on`), and unclear (e.g. `and`) relations, as pointed out by Chen *et al*. [1]. The problematic relation schemes can bring difficulties to denoising distant supervision.

Results on 1,700 Visual Genome Predicates. Visual distant supervision can alleviate the long-tail problem and therefore enables large-scale visual relation extraction. To

	Models	Accuracy			Mean Accuracy			# Non-zero Predicates		
		top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10
FS	Motif [8]	64.05	81.35	85.05	1.51	5.15	7.32	103	169	212
SS	Motif + DNS (Ours)	**66.10**	**84.26**	**87.72**	**3.08**	**9.49**	**13.44**	**218**	**362**	**435**

Table 2. Results of visual distant supervision on Visual Genome 1,700 predicates (%). FS: fully supervised, SS: semi-supervised.

investigate the effectiveness of visual distant supervision in handling large-scale visual relations, we refine the full Visual Genome predicates following the principles proposed by [1], resulting in $1,700$ well-defined predicates. In addition to top-K accuracy, to better focus on the long-tail performance, we also report top-K mean accuracy and the number of non-zero predicates (i.e., predicates with at least one correctly predicted instance). From the experimental results in Table 2, we observe that our model significantly outperforms its fully supervised counterpart. Notably, our model nearly doubles the top-K mean accuracy and the number of non-zero predicates, demonstrating the promising potential of visual distant supervision in handling large-scale visual relations in the future.

References

[1] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of CVPR*, pages 2580–2590, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186. Association for Computational Linguistics, 2019.

[3] Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, and Larry S Davis. Learning from noisy anchors for one-stage object detection. In *Proceedings of CVPR*, pages 10588–10597, 2020.

[4] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of CVPR*, pages 2117–2125, 2017.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of ECCV*, pages 740–755. Springer, 2014.

[6] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of CVPR*, pages 3716–3725, 2020.

[7] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of CVPR*, pages 1492–1500, 2017.

[8] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of CVPR*, pages 5831–5840, 2018.

Figure 2. Qualitative examples of model predictions in predicate classification task. We show top 10 predictions from (1) models that do not utilize human-labeled data, including weakly supervised and distantly supervised model, and (2) models that leverage human-labeled data, including fully supervised model and our semi-supervised model. Green edges: predictions that match Visual Genome annotations, blue edges: plausible predictions that are not labeled in Visual Genome, red edges: implausible predictions.