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: $\Lambda$: commonsense knowledge base
Require: $D_S$: distantly labeled image data
Require: $f(\cdot;\theta)$: any scene graph model, with parameter $\theta$
Optional: $\Phi$: external semantic signal
Randomly initialize the model parameter $\theta^0$
// Initial E step: estimate the probabilistic distant labels
Obtain distant labels $\mathbf{d} = \Psi(s, o, \Lambda)$
if external signal $\Phi$ 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 $\Phi$ 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: $\Lambda$: 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 $\theta$
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

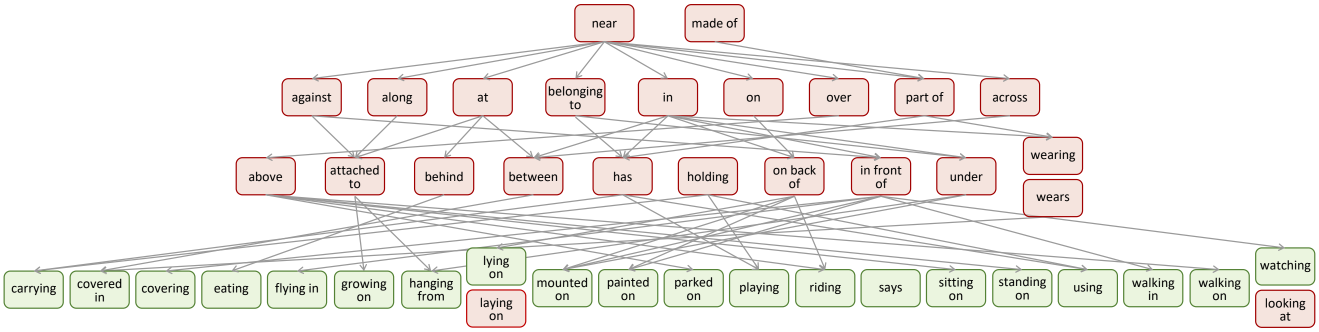

Figure 1. Dependencies of Visual Genome relations from Chen *et al.* [1]. Directed arrows: hypernyms. Stacked nodes: synonyms. Red nodes: removed relations. Green nodes: retained relations.

	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
	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
	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
	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.

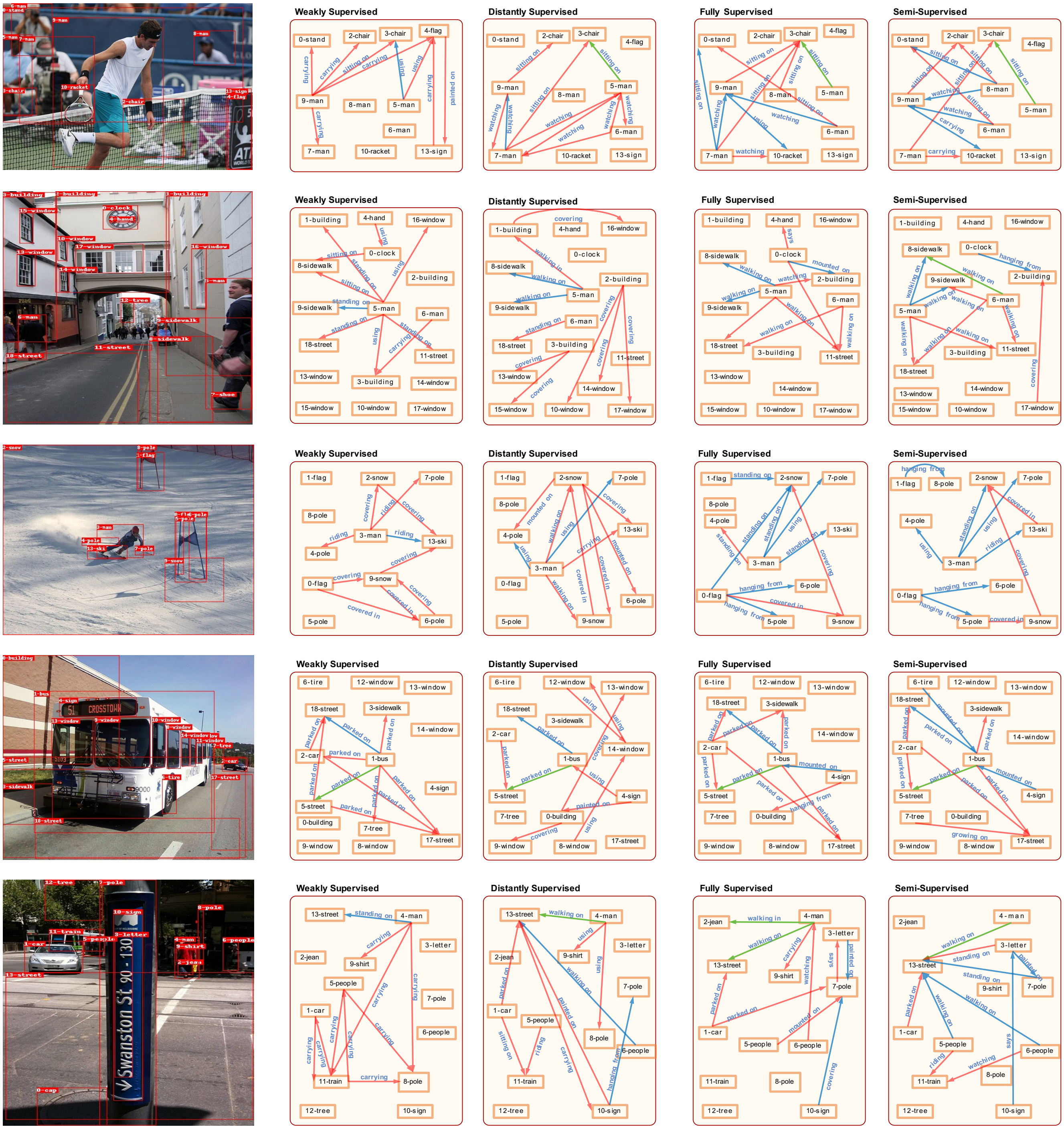

Figure 2. Qualitative examples of model predictions in predicate classification task. We show top 10 predictions from (1) models that do not utilize human-labeled data, including weakly supervised and distantly supervised model, and (2) models that leverage human-labeled data, including fully supervised 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.