Graphical Contrastive Loss for Scene Graph Parsing Supplementary Material

Ji Zhang[1,2], Kevin J. Shih[2], Ahmed Elgammal[1], Andrew Tao[2], Bryan Catanzaro[2]

[1]Department of Computer Science, Rutgers University
[2]Nvidia Corporation

1. Complexity Analysis for the Losses

Computational analysis for our sampling and inference procedures are provided below. We look at the case where the subject s_i is fixed and we vary object for positive/negative pairings. The reverse case (object fixed, subject varies) has the same complexity. All sampling is conducted on the entities of a single image per batch. The set of entities include ground truth bounding boxes, as well as any detector output with > 0.5 IOU to ground truth entities.

For the Class Agnostic Loss L_1, the computational complexity of the sampling procedure is $O(N^2)$, where N is the upper bounded on number of sampled entities per image. In practice, for each subject, we randomly sample at most K non-related objects (negative pairings), which makes the actual complexity $O(NK)$.

For the Entity Class Aware Loss L_2, the sampling procedure is the same as with L_1, except that we need to keep only those non-related objects that are of class c, *i.e.*, the object class of the current o in the sampled (s, o) pair. This involves a filtering operation on the K objects which takes $O(K)$ time, therefore the overall complexity is still $O(NK)$.

The analysis for the Predicate Class Aware Loss L_3 is similar to that of L_2, except that the filtering operation looks at the predicate class e instead of the object class c. The overall complexity is also $O(NK)$.

We set $N = 512$ and $K = 64$ per batch in practice.

2. Full Results on VG and VRD

We present full experimental results compared with all previous competitive methods on Visual Genome (VG) and Visual Relation Detection (VRD) datasets in Table 1 and Table 2. We also show results of the baseline RelDN without our Graphical Contrastive Losses (L_0 only).

On VG, we observe that our losses achieve smaller gains over cross-entropy loss than it does on OpenImages_mini (Table 1,2 in the main paper). The reasons are two-fold: 1) One of the few dominant relationship types in VG is possessive, *e.g.*, "ear of man", which has much less entity confusion issues; 2) The $Recall@k$ metric is less strict than mAP. If there is an image with only one ground truth, then Recall@100 will always be 100% as long as this ground truth target is within the top 100 model predictions, regardless of the ranking of the 100 outputs. As such, the small improvements in ranking the top 100 will not affect the score. Nevertheless, the improvements by our loss is still non-trivial and consistent on all metrics under different values of k.

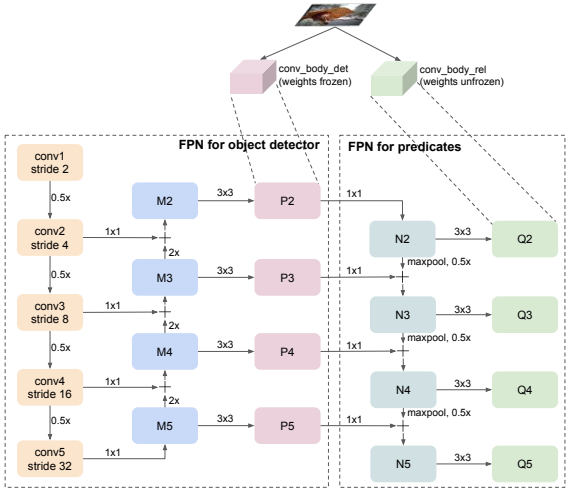


Figure 1: The lateral connections for predicate CNN features. P2-P5 are fixed and used as entity detector features, while Q2-Q5 are predicate features, which are trained to transform entity features to features for predicates.

For the interest of future work, we also show results using a better backbone, ResNeXt-101-FPN [8, 4] for the entity detector in Table 1.

On VRD, the gap between L_0 *only* and the full model is smaller when pre-trained on ImageNet than on COCO detection. We believe the stronger localization features from pre-training on COCO is much easier for our proposed model and losses to leverage.

	Graph Constraint									No Graph Constraint					
	SGDET			SGCLS			PRDCLS			SGDET		SGCLS		PRDCLS	
Recall at	20	50	100	20	50	100	20	50	100	50	100	50	100	50	100
VRD[5]	-	0.3	0.5	-	11.8	14.1	-	27.9	35.0	-	-	-	-	-	-
Associative Embedding[6]	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4	9.7	11.3	26.5	30.0	68.0	75.2
Message Passing[9]	-	3.4	4.2	-	21.7	24.4	-	44.8	53.0	-	-	-	-	-	-
Message Passing+[12]	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3	22.0	27.4	43.4	47.2	75.2	83.6
Frequency[12]	17.7	23.5	27.6	27.7	32.4	34.0	49.4	59.9	64.1	25.3	30.9	40.5	43.7	71.3	81.2
Frequency+Overlap[12]	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2	28.6	34.4	39.0	43.4	75.7	82.9
MotifNet-NOCONTEXT[12]	21.0	26.2	29.0	31.9	34.8	35.5	57.0	63.7	65.6	29.8	34.7	43.4	46.6	78.8	85.9
MotifNet-LeftRight[12]	**21.4**	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1	**30.5**	35.8	44.5	47.7	81.1	88.3
RelDN, L_0 only	20.8	28.1	32.5	**36.1**	36.7	36.7	66.7	68.3	68.3	30.1	36.4	**48.9**	**50.8**	93.7	97.7
RelDN	21.1	**28.3**	**32.7**	**36.1**	**36.8**	**36.8**	**66.9**	**68.4**	**68.4**	30.4	**36.7**	**48.9**	**50.8**	**93.8**	**97.8**
RelDN (X-101-FPN)	22.5	31.0	36.7	38.2	38.9	38.9	67.2	68.7	68.8	32.6	40.0	51.7	53.6	94.0	97.8

Table 1: Comparison with state-of-the-arts on VG. L_0 **only** is the RelDN without our losses. We also include results of our model with ResNeXt-101-FPN as the backbone for future work reference.

	Relationship		Phrase		Relationship Detection						Phrase Detection					
	free k				k = 1		k = 10		k = 70		k = 1		k = 10		k = 70	
Recall at	50	100	50	100	50	100	50	100	50	100	50	100	50	100	50	100
PPRFCN*[13]	14.41	15.72	19.62	23.75	-	-	-	-	-	-	-	-	-	-	-	-
VTransE*	14.07	15.20	19.42	22.42	-	-	-	-	-	-	-	-	-	-	-	-
SA-Full*[7]	15.80	17.10	17.90	19.50	-	-	-	-	-	-	-	-	-	-	-	-
DR-Net*[1]	17.73	20.88	19.93	23.45	-	-	-	-	-	-	-	-	-	-	-	-
ViP-CNN[2]	17.32	20.01	22.78	27.91	17.32	20.01	-	-	-	-	22.78	27.91	-	-	-	-
VRL[3]	18.19	20.79	21.37	22.60	18.19	20.79	-	-	-	-	21.37	22.60	-	-	-	-
CAI*[14]	20.14	23.39	23.88	25.26	-	-	-	-	-	-	-	-	-	-	-	-
KL distilation[11]	22.68	31.89	26.47	29.76	19.17	21.34	22.56	29.89	22.68	31.89	23.14	24.03	26.47	29.76	26.32	29.43
Zoom-Net[10]	21.37	27.30	29.05	37.34	18.92	21.41	-	-	21.37	27.30	24.82	28.09	-	-	29.05	37.34
CAI + SCA-M[10]	22.34	28.52	29.64	38.39	19.54	22.39	-	-	22.34	28.52	25.21	28.89	-	-	29.64	38.39
RelDN, L_0 only (ImageNet)	21.62	26.12	28.59	35.18	19.57	22.61	21.62	26.12	21.62	26.12	26.39	31.28	28.59	35.18	28.59	35.18
RelDN (ImageNet)	21.52	26.38	28.24	35.44	19.82	22.96	21.52	26.38	21.52	26.38	26.37	31.42	28.24	35.44	28.24	35.44
RelDN, L_0 only (COCO)	26.67	32.55	33.29	41.25	24.30	27.91	26.67	32.55	26.67	32.55	31.09	**36.42**	33.29	41.25	33.29	41.25
RelDN (COCO)	**28.15**	**33.91**	**34.45**	**42.12**	**25.29**	**28.62**	**28.15**	**33.91**	**28.15**	**33.91**	**31.34**	**36.42**	**34.45**	**42.12**	**34.45**	**42.12**

Table 2: Comparison with state-of-the-art on VRD (− means unavailable / unknown). Same with Table 1, L_0 **only** is the RelDN without our losses. "Free k" means considering k as a hyper-parameter that can be cross-validated.

L_0	L_1	L_2	L_3	R@50	mAP$_{rel}$	mAP$_{phr}$	score	mAP$^*_{rel}$	mAP$^*_{phr}$	score*
✓				74.67	35.28	41.04	45.46	33.87	38.99	44.08
✓	✓			75.06	**44.18**	**50.19**	**52.76**	35.24	40.30	45.23
✓		✓		74.64	36.19	41.71	46.09	34.67	39.61	44.64
✓			✓	74.88	34.80	40.47	45.08	34.92	40.01	44.95
✓	✓	✓		75.03	35.10	41.18	45.52	35.09	40.22	45.13
✓	✓		✓	**75.30**	43.96	49.61	52.49	34.89	39.87	44.96
✓		✓	✓	75.00	35.83	41.32	45.86	34.62	39.70	44.73
✓	✓	✓	✓	74.94	39.09	44.47	48.41	**35.82**	**40.43**	**45.49**

Table 3: Ablation Study on our losses with the official mAP$_{rel}$, mAP$_{phr}$ and score metrics. Metric marked with a * means the predicate "under" and "hits" are excluded from evaluation. The fluctuating numbers in mAP$_{rel}$, mAP$_{phr}$ and score indicate that the mAP metrics are unstable and unreliable, while when "under" and "hits" are excluded, all the results become consistent with Table 1 in the main paper.

3. Results Under the Official mAP metrics

In our main paper, we use a class-frequency weighted mAP (*wmAP*) for model comparison, with the aim of de-emphasizing the classes with only a handful of test examples (specifically "under" and "hits"). This is because their small sample size resulted in extremely large variances between runs. Here, we show our ablation studies using the official uniform-class-weighting evaluation metrics, *mAP*$_{rel}$, *mAP*$_{phr}$ and *score*, as defined in Section 6.1 in the main paper. We also include *mAP*$^*_{rel}$, *mAP*$^*_{phr}$ and *score**, which is the standard mAP and score excluding "under" and "hits" in the evaluation. Table 3 presents ablation study results on loss components, corresponding to Table 1 in the paper. Table 4 shows comparison between the L_0-only model against the model with our losses on the 100 selected images, corresponding to Table 2 in the paper. In Table 3 the variation of numbers using mAP and score demonstrates the necessity of de-emphasizing the extremely infrequent classes. Note that the mAP*-based columns show a similar trend to our wmAP-based results from the paper. In Table 4, the model with our losses is still better than the L_0-only model by a non-trivial margin, mainly because the former outperform the latter on almost every per-class AP metric for those 5 selected classes. Note that since "under" and "hits" are not in the 100 image subset, there is no need to evaluate with *mAP*$^*_{rel}$, *mAP*$^*_{phr}$ and *score**.

4. An Alternative for the Predicate CNN

We want to answer a natural question about the predicate CNN branch: can we use a less expensive feature extractor for predicates instead of a full CNN branch? We follow the idea of FPN [4] and add laterally connected layers to

	R@50	mAP$_{rel}$	mAP$_{phr}$	score
L_0	61.72	25.20	35.37	36.57
$L_0 + L_1 + L_2 + L_3$	**62.65**	**26.77**	**36.79**	**37.95**

Table 4: Comparison of our model with Graphical Contrastive Loss *vs*. without the loss on 100 images containing the 5 classes that suffer from the two aforementioned confusions, selected via visual inspection on a random set of images. The metrics are the official mAP$_{rel}$, mAP$_{phr}$ and the score. The "under" and "hits" predicates are not in this 100 image subset.

	R@50	wmAP$_{rel}$	wmAP$_{phr}$	score$_{wtd}$
ImageNet init	74.82	34.93	37.96	44.12
entity detector	74.85	35.06	38.15	44.25
obj transform	**75.03**	35.21	38.12	44.34
fully trained	74.94	**35.54**	**38.52**	**44.61**

Table 5: Predicate branch comparison on OI_mini. "obj transform" means using the lateral connected layers as the predicate feature extractor. All other abbreviations are the same with Table 5 in the main paper.

the entity detector's CNN layers, which are trained to transform entity features to predicate-relevant features. Figure 1 illustrates these layers.

Table 5 shows that using lateral connections to transform entity features is better than using fixed entity features, but still inferior than the separate predicate CNN, which demonstrates necessity of the latter.

5. Qualitative Results

In Figure 2 we provide four example images where our losses correct the false predictions made by the L_0 only model. Both the Entity Instance Confusion and the Proximal Relationship Ambiguity issues are included here. In the fourth row, the L_0 only model is confused between two entity instances, *i.e*., which person is holding the microphone, while our losses manage to refer to the correct one. In the third row the relationship between the guitar player and the drum is ambiguous. Here, the L_0 only model fails by predicting a false-positive, but our model trained with all losses correctly detects no relationship there.

6. Examples of the 100 Image Subset

Figure 3 shows several examples, randomly selected from the 100 image subset that we use to demonstrate the advantage of our losses (described in Section 6.2 of the main paper). These images contain very challenging relationships such as two women holding two cellphones while sitting very closely to each other, or three men interacting with (riding on) three horses where two of them are occluded since they are very close.

References

[1] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 2

[2] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. In *CVPR*, 2017. 2

[3] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 2017. 2

[4] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1, 2

[5] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 2

[6] A. Newell and J. Deng. Pixels to graphs by associative embedding. In *NIPS*, 2017. 2

[7] J. Peyre, I. Laptev, C. Schmid, and J. Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017. 2

[8] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 1

[9] D. Xu, Y. Zhu, C. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 2

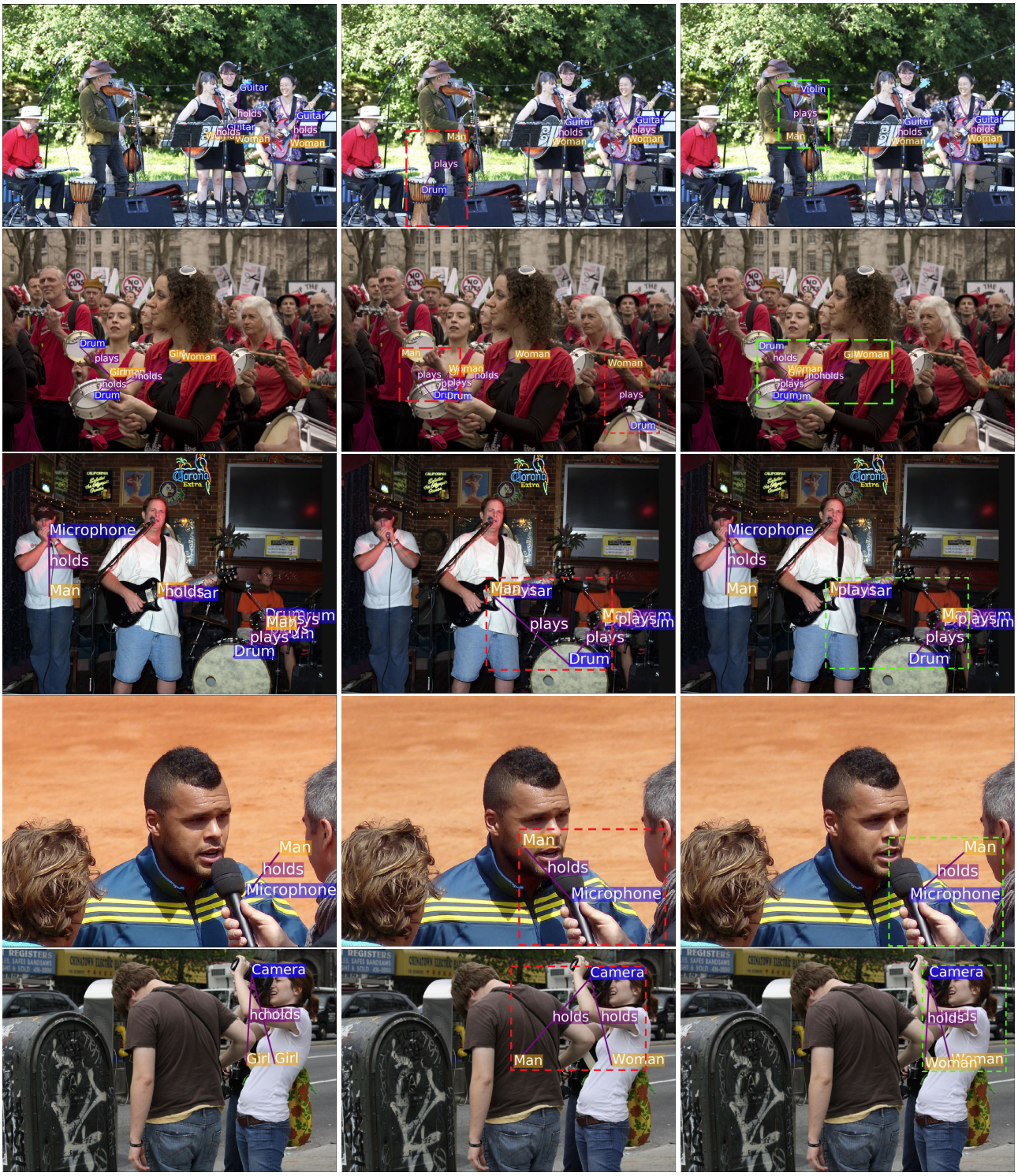
[10] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*, 2018. 2

[11] R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, 2017. 2

[12] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 2

[13] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *CVPR*, 2017. 2

[14] B. Zhuang, L. Liu, C. Shen, and I. Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, 2017. 2



(a) ground truth (b) L_0 only (c) all losses

Figure 2: Example images where RelDN with only L_0 predicts incorrectly while our loss succeeds. For each image we check the number of its ground truth relationships, then we output the same number of top predictions from a model to see its ranking accuracy. Red boxes in (b) highlight the false predictions from RelDN with L_0 only and green boxes in (c) highlight the correct ones from RelDN with all losses.

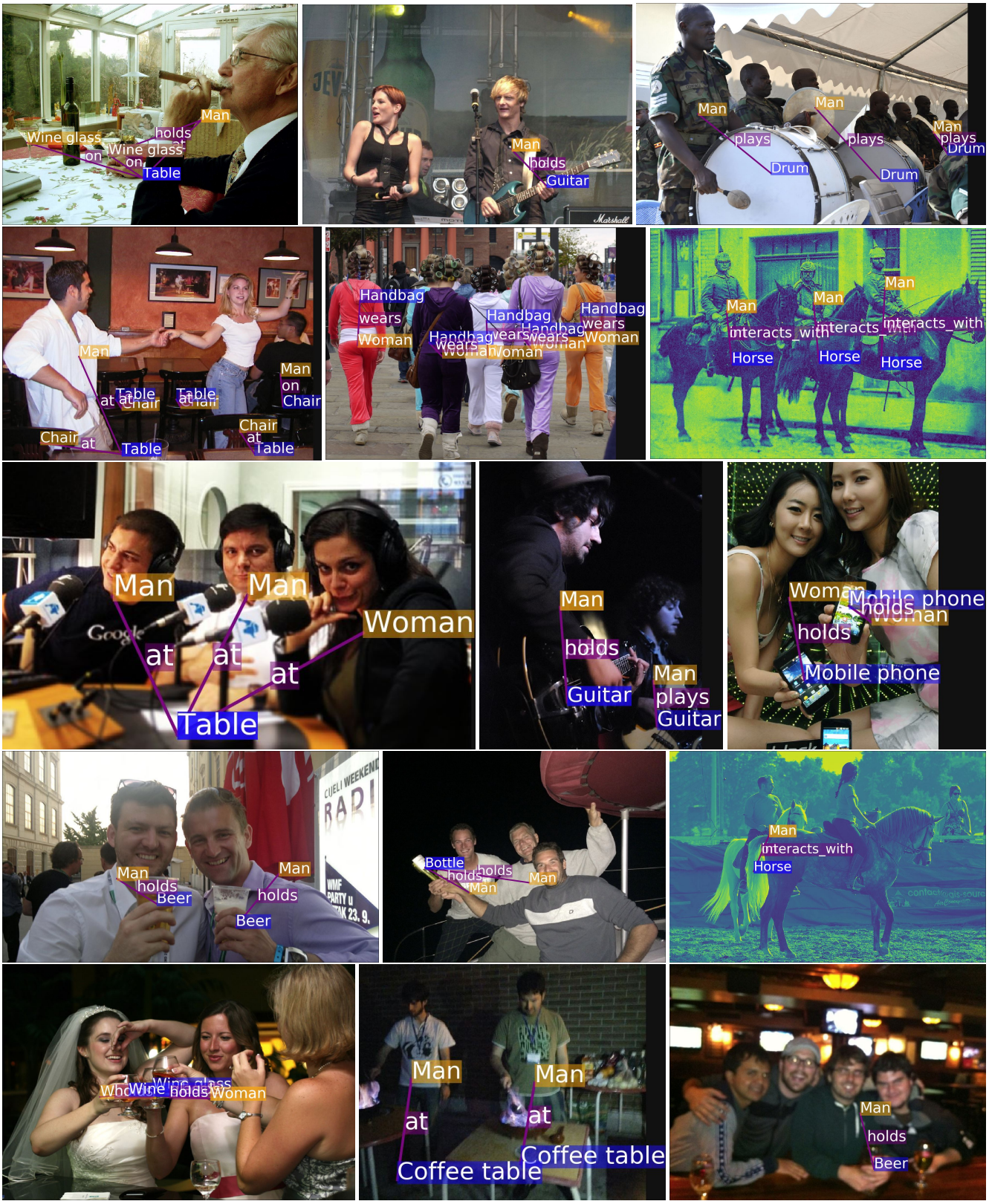


Figure 3: Example images of the 100 image subset with ground truth relationships. The subset contains five predicates where the Entity Instance Confusion and Proximal Relationship Ambiguity commonly occur.