



Improving Visual Relation Detection using Depth Maps

Sahand Sharifzadeh

Sina M. Baharlou

Max Berrendorf

Rajat Koner

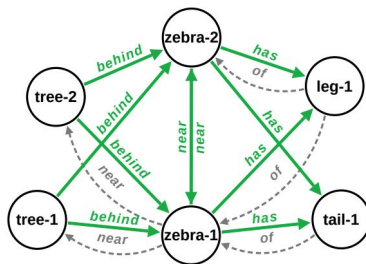
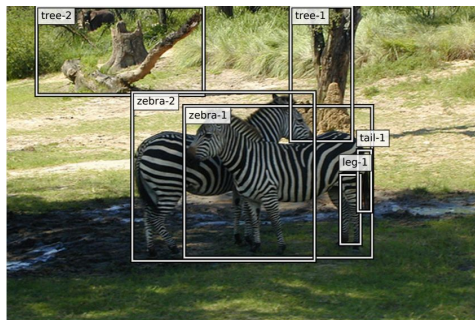
Volker Tresp

Ludwig Maximilian University of Munich



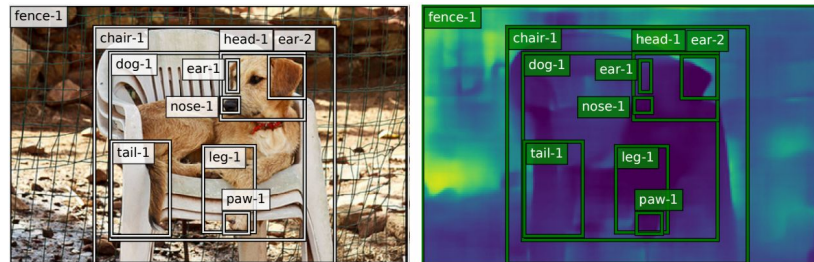
Visual Relation Detection

- Detecting relations between objects in an image



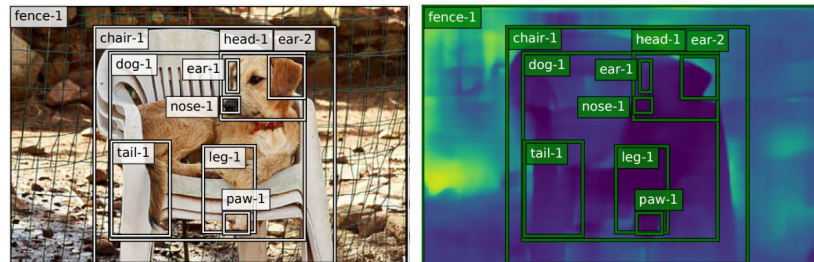
Depth Maps

1. We argue that depth maps can additionally provide valuable information about an object's relations as they provide the object's distance from the camera.



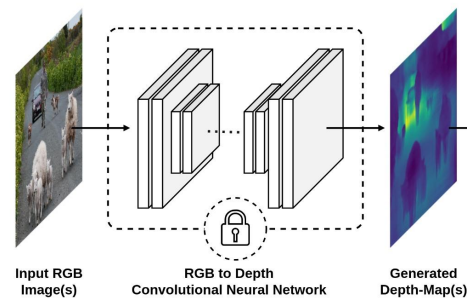
Depth Maps

1. We argue that depth maps can additionally provide valuable information about an object's relations as they provide the object's distance from the camera.
2. Unfortunately, most available image datasets do not provide depth maps, because the acquisition of depth maps is a cumbersome task requiring specialized hardware.



Synthetic Depth Maps

We tackle this issue by synthetically generating [1] the corresponding pseudo depth maps from 2D images of Visual Genome [2].



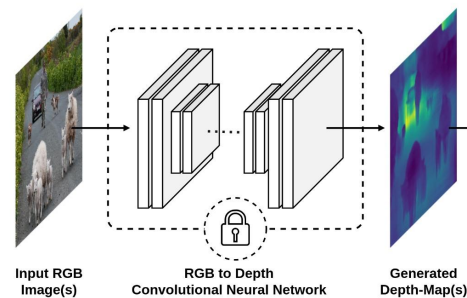
[1] Laina, Iro, et al. "Deeper depth prediction with fully convolutional residual networks." 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016.

[2] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123.1 (2017): 32-73.

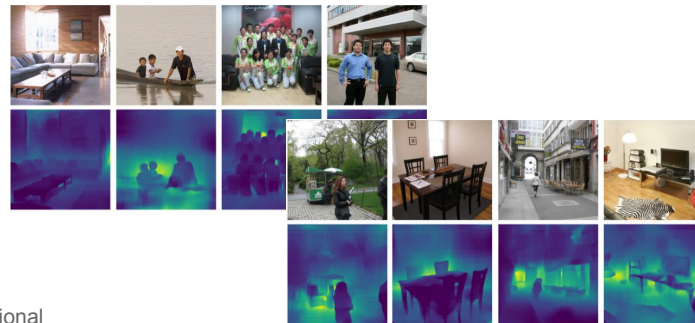
[3] <https://github.com/Sina-Baharlou/Depth-VRD>

Synthetic Depth Maps

We tackle this issue by synthetically generating [1] the corresponding pseudo depth maps from 2D images of Visual Genome [2].



We release the generated depth maps as a separate dataset called *VG-Depth* [3].



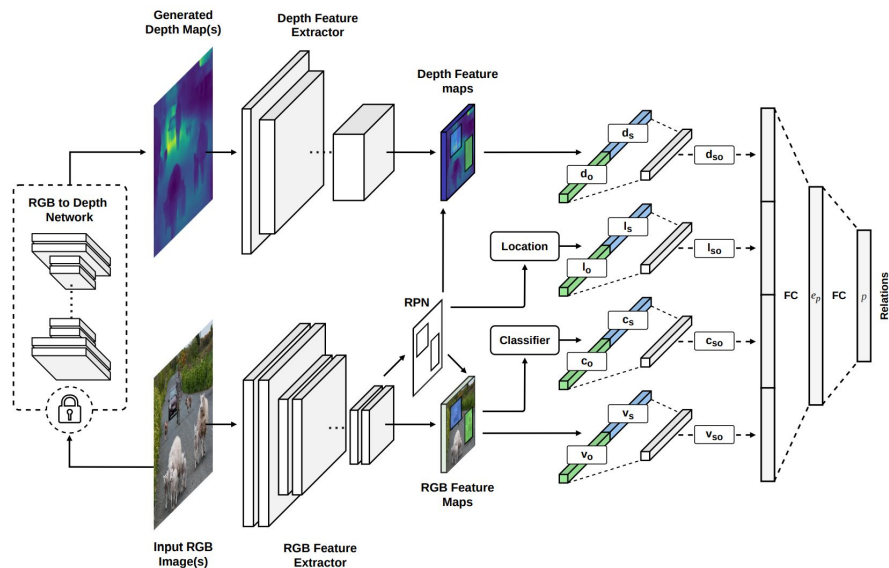
[1] Laina, Iro, et al. "Deeper depth prediction with fully convolutional residual networks." 2016 Fourth international conference on 3D vision (3DV). IEEE, 2016.

[2] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123.1 (2017): 32-73.

[3] <https://github.com/Sina-Baharlou/Depth-VRD>

Multimodal Architecture

The object information extracted from depth maps and RGB images, i.e. class labels, location vectors, RGB and depth features, are the basis for relation detection in our simple yet effective framework.






Experiments

- We test our approach on the Visual Genome [1].
- Metrics:
 - [Micro] Recall@K

[1] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123.1 (2017): 32-73.



Experiments

- We test our approach on the Visual Genome [1].
- Metrics:
 - [Micro] Recall@K
 - **Macro Recall@K:**

$$\text{MACRO RECALL@K} = \sum_{(s,p,o) \in \mathcal{T}_p} \frac{\text{MICRO R@K}(p)}{|\mathcal{T}_p|}$$

Experiments

- We test our approach on the Visual Genome [1].
- Metrics:
 - [Micro] Recall@K
 - **Macro Recall@K:**

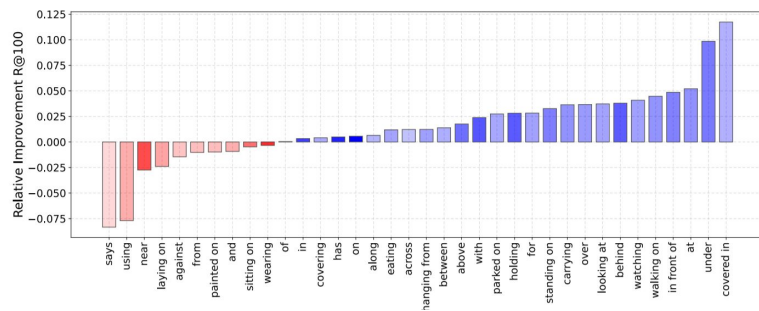
$$\text{MACRO RECALL@K} = \sum_{(s,p,o) \in \mathcal{T}_p} \frac{\text{MICRO R@K}(p)}{|\mathcal{T}_p|}$$

Strategy Task Metric	Macro Predicate Pred.			Micro Predicate Pred.		
	R@100	R@50	R@20	R@100	R@50	R@20
models	VTransE [27]	-	-	-	62.87	62.63
	Yu's-S [15]	-	-	-	49.88	-
	Yu's-S+T [15]	-	-	-	55.89	-
	IMP [16]	-	-	-	53.00	44.80
	Graph R-CNN [18]	-	-	-	59.10	54.20
	NM [17]	14.39	13.20	10.25	67.10	65.20
ablations	Ours - <i>d</i>	9.51	8.46	6.35	54.72	51.90
	Ours - <i>c</i>	15.65	13.09	8.56	64.82	60.54
	Ours - <i>v</i>	13.88	12.24	8.99	61.72	58.50
	Ours - <i>l</i>	5.19	4.66	3.57	49.07	46.13
	Ours - <i>v, d</i>	15.47	14.04	10.83	62.88	60.52
	Ours - <i>l, v, d</i>	15.76	14.40	11.07	63.06	60.83
	Ours - <i>l, c, d</i>	21.67	19.56	15.12	67.97	66.09
	Ours - <i>l, c, v</i>	19.16	17.72	13.93	67.94	66.06
	Ours - <i>l, c, v, d</i>	22.72	20.74	16.40	68.00	66.18
					59.44	

[1] Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." International journal of computer vision 123.1 (2017): 32-73.

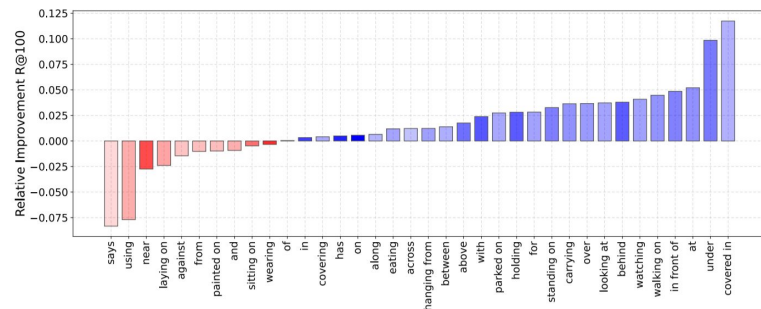
Experiments

- Per Predicate Improvement

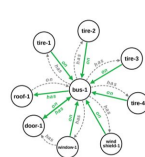
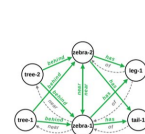
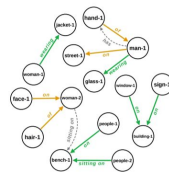
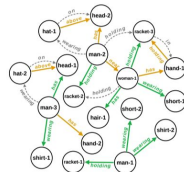
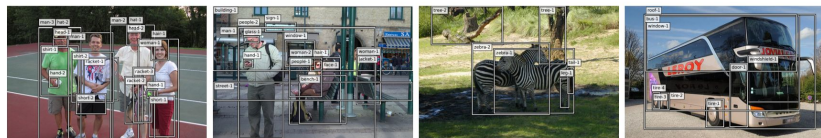


Experiments

- Per Predicate Improvement



- Qualitative Results





Summary

1. We perform an extensive study on the effect of using different sources of object information in visual relation detection. We show in our empirical evaluations using the VG dataset, that our model can outperform competing methods by a margin of up to 8% points.
 2. We release a new synthetic dataset VG-Depth, to compensate for the lack of depth maps in Visual Genome.
 3. We propose Macro Recall@K as a competitive metric for evaluating the visual relation detection performance in highly imbalanced datasets such as Visual Genome.
- 