Supplemental Materials

Anonymous CVPR submission

Paper ID 1142

1. Proof of Equation 4 in Section 4.2

The posterior distribution of r is given by

$$p(r|s, o, \mathbf{x}_r; \mathbf{W}) \propto \exp(\psi_r(r|\mathbf{x}_r; \mathbf{W}_r)+ \phi_{rs}(r, s|\mathbf{W}_{rs}) + \phi_{ro}(r, o|\mathbf{W}_{ro})). \quad (1)$$

Here, we have

1. The unary potential $\psi_r(r|\mathbf{x}_r)$ is assumed to be a linear functional of $\mathbf{x}_r$ for each predicate r, then we can write $\psi_r(r|\mathbf{x}_r) := \mathbf{a}_r^T \mathbf{x}_r$. Combining the linear functionals for all categories, we can form a coefficient matrix $\mathbf{W}_r = [\mathbf{a}_{r_1}^T, \mathbf{a}_{r_2}^T, ..., \mathbf{a}_{r_{|\mathcal{R}|}}^T]$. Thus, $\psi_r(r|\mathbf{x}_r; \mathbf{W}_r) = \mathbf{1}_r^T \mathbf{W}_r \mathbf{x}_r$.

2. Both r and s be categorical variables. Hence, the potential ϕ_{rs} can be represented by a matrix of size $|\mathcal{R}| \times |\mathcal{O}|$, where $\mathcal{R}$ is the set of all relationship predicates while $\mathcal{O}$ is the set of all object categories. Particularly, let $\mathbf{1}_r$ and $\mathbf{1}_s$ be indicator vectors for r and s, then we have $\phi_{rs}(r, s|\mathbf{W}_{rs}) = \mathbf{1}_r^T \mathbf{W}_{rs} \mathbf{1}_s$.

3. Likewise, the potential ϕ_{ro} can also be characterized by a matrix $\mathbf{W}_{ro}$, such that $\phi_{ro}(r, o|\mathbf{W}_{ro}) = \mathbf{1}_r^T \mathbf{W}_{ro} \mathbf{1}_o$.

Let $\mathbf{q}_r(r) = p(r|s, o, \mathbf{x}_r; \mathbf{W})$, then Eq.(1) can be rewritten:

$$\mathbf{q}_r(r) \propto \exp\left(\mathbf{1}_r^T \mathbf{W}_r \mathbf{x}_r + \mathbf{1}_r^T \mathbf{W}_{rs} \mathbf{1}_s + \mathbf{1}_r^T \mathbf{W}_{ro} \mathbf{1}_o\right) \quad (2)$$
$$= \exp\left(\mathbf{1}_r^T \left(\mathbf{W}_r \mathbf{x}_r + \mathbf{W}_{rs} \mathbf{1}_s + \mathbf{W}_{ro} \mathbf{1}_o\right)\right). \quad (3)$$

This equation can be interpreted as follows. The expression $\mathbf{e} = \mathbf{W}_r \mathbf{x}_r + \mathbf{W}_{rs} \mathbf{1}_s + \mathbf{W}_{ro} \mathbf{1}_o$ is a vector of length $|\mathcal{R}|$, and the operator $\mathbf{1}_r^T \mathbf{e}$ takes the r-th entry. $\mathbf{q}_r$ is comprised of the normalized exponents of these entries, and thus can be written as

$$\mathbf{q}_r = \sigma(\mathbf{W}_r \mathbf{x}_r + \mathbf{W}_{rs} \mathbf{1}_s + \mathbf{W}_{ro} \mathbf{1}_o) \quad (4)$$

Here, σ is the softmax function that produces a vector of normalized exponents. This completes the proof.

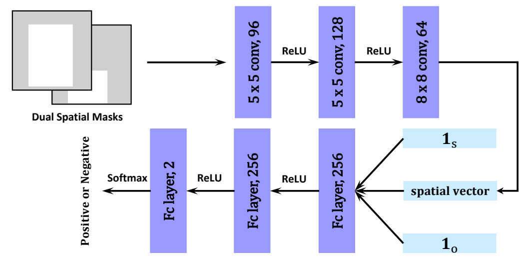


Figure 1: The network for pair filtering.

2. Pair Filter

As mentioned in the paper, we use a simple network to filter out part of the pairs before feeding them to the main DR-Net for further analysis. Here are some technical details about the network. Figure 1 shows the architecture of this network. The network comprises three convolutional layers followed by three fully-connected layers. These layers are interleaved with *ReLU* activations. It is designed to be relatively shallow, so that it can perform the filtering with low cost. To train this network, we randomly sample pairs of bounding boxes from each training image, treating those with 0.5 IoU (or above) with any ground-truth pairs as positive samples, and the rest as negative samples.

In testing, from n detected objects, we can form $n(n-1)$ pairs. We use this filter to remove 40% of them, retaining 60%. This filtering rate was chosen empirically based on the overall empirical performance on a validation set.

3. More Examples

The following tables show more examples of our results. As discussed in the experiment section, the annotations in data sets are not complete. Some true relationships are missing in the data sets.

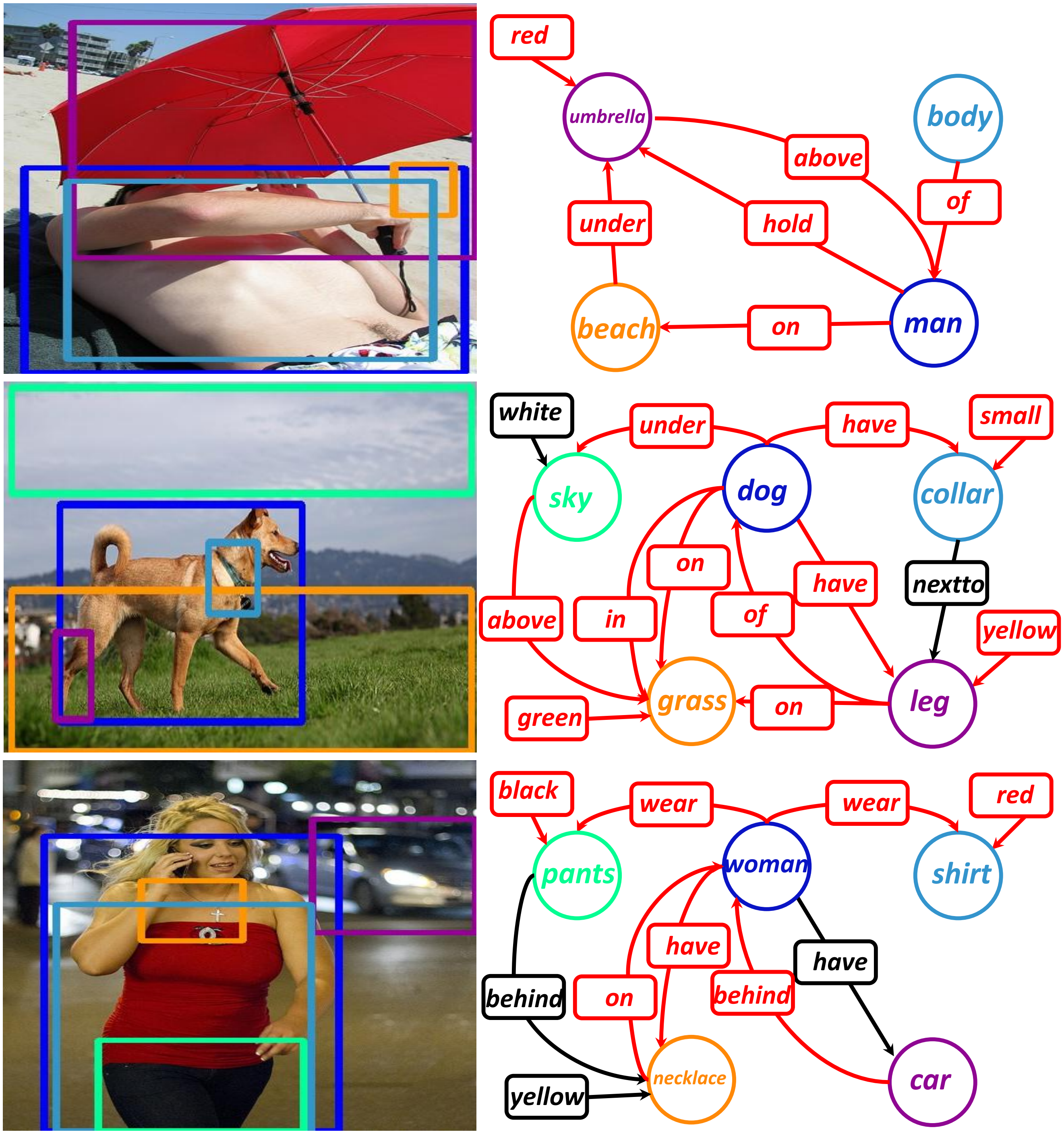


Table 1: This table lists more samples of scene graph generation, where red edges indicate correct prediction, and black edges indicate wrong prediction.

Image	Ground truth relationship triplets	Top detections not in ground truth	
	(hat, on, person) (glasses, on, person) (person, wear, jacket) (person, wear, hat) (pseron, wear, glasses) (person, stand behind, person)	(jacket, on, person) (sky, above, person) (sky, above, glasses) (jeans, on, person) (person, wear, jeans)	0.80 0.49 0.46 0.46 0.44
	(person, wear, hat) (person, next to, truck) (person, wear, shirt) (person, front of, traffic light) (traffic light, behind, person) (truck, next to, person)	(hat, on, truck) (hat, on, person) (jacket, on, truck) (truck, has, hat) (jacket, on, person)	0.57 0.53 0.52 0.50 0.49
	(person, on, skateboard) (person, wear, shirt) (skateboard, under, person)	(shirt, on, person) (shirt, on, skateboard) (shirt, on, box) (skateboard, under, person) (skateboard, under, skirt)	0.53 0.53 0.36 0.26 0.21
	(airplane, has, wheel) (airplane, on, street) (wheel, below, airplane)	(wheel, on, street) (street, has, wheel) (luggage, on, street) (street, under, airplane) (wheel, on, airplane)	0.86 0.77 0.52 0.51 0.38

Table 2: Examples of visual relationship detection results. **Col 1:** Input images. **Col 2:** Ground-truth triplets. Red indicates those correctly recalled in the top-50 list. **Col 3:** Top-5 predictions for each image. Blue indicates predictions that are actually correct but are not included in the annotations. We can observe quite reasonable predictions from the proposed detector. It can often successfully detect correct relationships that are not annotated.

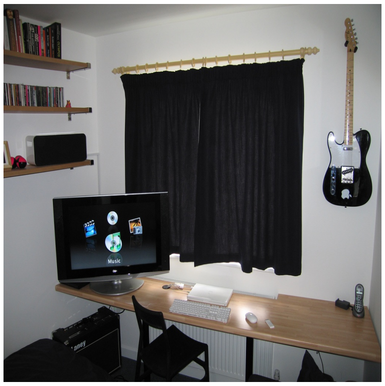
Image	Ground truth relationship triplets	Top detections not in ground truth	
	(monitor, on, desk) (chair, front of, desk) (phone, on, desk) (keyboard, on, desk) (mouse, on, desk)	(paper, on, desk) (desk, under, monitor) (desk, under, mouse) (box, on, desk) (desk, under, phone)	0.52 0.42 0.37 0.33 0.31
	(person, has, jacket) (person, has, pants) (person, wear, shirt) (person, behind, person)	(pants, on, person) (shirt, on, person) (person, wear, pants) (shirt, above, pants) (shirt, on, phone)	0.62 0.62 0.53 0.45 0.39
	(monitor, next to, keyboard) (keyboard, next to, mouse) (keyboard, next to, monitor) (keyboard, front of, monitor) (mouse, next to, keyboard) (person, next to, monitor) (monitor, next to, person) (monitor, behind, keyboard)	(monitor, above, trash can) (trash can, under, monitor) (monitor, above, keyboard) (keyboard, front of, monitor) (person, has, keyboard)	0.29 0.25 0.25 0.17 0.15
	(person, wear, shirt) (person, wear, shorts) (person, wear, shoes) (person, has, ball) (shorts, above, shoes)	(shoes, on, person) (sky, above, person) (shorts, on, person) (shorts, on, ball) (ball, above, shoes)	0.50 0.49 0.48 0.48 0.47

Table 3: Continue Table 2 – more detection results.