University of Pittsburgh
School of Computing and Information

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Context for Object Detection via Lightweight Global and Mid-level Representations

Mesut Erhan Unal and **Adriana Kovashka**
University of Pittsburgh

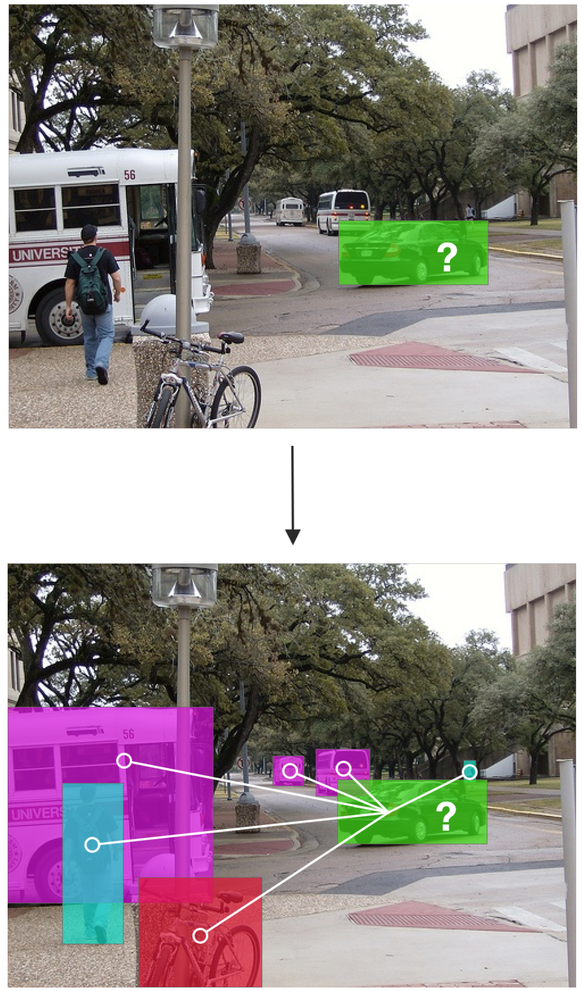

Motivation

- Context is an important mechanism that makes visual recognition easier for humans [Palmer, 1975; Biederman et al., 1982], hence it is natural to also model context in machine perception.

- State-of-the-art two-stage object detection frameworks (e.g. Faster R-CNN) perform classification and localization tasks for each region in isolation, ignoring what is in the rest of the image.

Motivation

- Prior work tries to model context in a manner which is expensive from both computational and human labeling point of view. For instance;
  - [Chen et al., 2018] requires densely-labeled datasets such as Visual Genome and ADE.
  - [Liu et al., 2018] employs recurrent units for belief-propagation.
- We propose a novel approach for context-aware object detection by employing a lightweight belief-propagation mechanism which operates on visual representations of regions and the scene, as well as the spatial relationships between regions.
- We also experiment with capturing similarities between regions at a semantic level by modeling class co-occurrence and linguistic similarity between class names.



Approach

Our work builds on top of Structure Inference Net **(SIN)**, proposed in [Liu et al., 2018]. In SIN, a message passed from region i to j is weighted by a constant $e_{i \to j}$, such that:

$$e_{i \to j} = \underbrace{ReLU(W_g R_{i \to j})}_{\text{Spatial relationship}} * \underbrace{tanh(W_v[f^i, f^j])}_{\text{Visual relationship}}$$

Approach

Our work builds on top of Structure Inference Net **(SIN)**, proposed in [Liu et al., 2018]. In SIN, a message passed from region i to j is weighted by a constant $e_{i \to j}$, such that:

$$e_{i \to j} = \underbrace{ReLU(W_g R_{i \to j})}_{\text{Spatial relationship}} * \underbrace{tanh(W_v[f^i, f^j])}_{\text{Visual relationship}} * \underbrace{\ldots\ldots}_{\text{Semantic relationships}}$$

- The first set of models we experiment with employ SIN's graph structure inference module as a post-processing step (after Faster R-CNN) to utilize semantic cues using class predictions.

Approach

Our work builds on top of Structure Inference Net **(SIN)**, proposed in [Liu et al., 2018]. In SIN, a message passed from region i to j is weighted by a constant $e_{i \to j}$, such that:

$$e_{i \to j} = \underbrace{ReLU(W_g R_{i \to j})}_{\text{Spatial relationship}} * \underbrace{tanh(W_v[f^i, f^j])}_{\text{Visual relationship}} * \underbrace{\ldots\ldots}_{\text{Semantic relationships}}$$

- The first set of models we experiment with employ SIN's graph structure inference module as a post-processing step (after Faster R-CNN) to utilize semantic cues using class predictions.

- Our second set of models replaces SIN's recurrent units with a lightweight mechanism for belief-propagation on region graph.

Approach

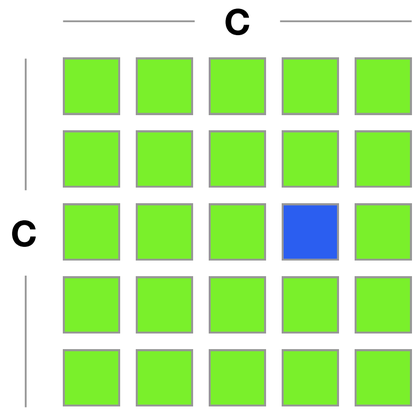
- **Base:** Models the co-occurrence of object categories based on the backbone detector’s best set of guesses to capture the semantic relationship between regions.

Approach

- **Base:** Models the co-occurrence of object categories based on the backbone detector’s best set of guesses to capture the semantic relationship between regions.
- **Scene:** Updates scene representation at the end of each message-passing round, then uses this new representation for the next round.

Approach

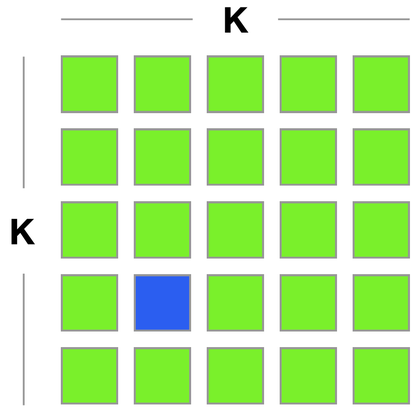
- **Base:** Models the co-occurrence of object categories based on the backbone detector's best set of guesses to capture the semantic relationship between regions.
- **Scene:** Updates scene representation at the end of each message-passing round, then uses this new representation for the next round.
- **Attr1:** Models mid-level semantic relationships between regions using object class attributes. Having built an attribute dictionary, this model learns a class-class similarity matrix over a latent attribute space and uses the highest class predictions to retrieve attribute similarity of regions.



C: Number of classes

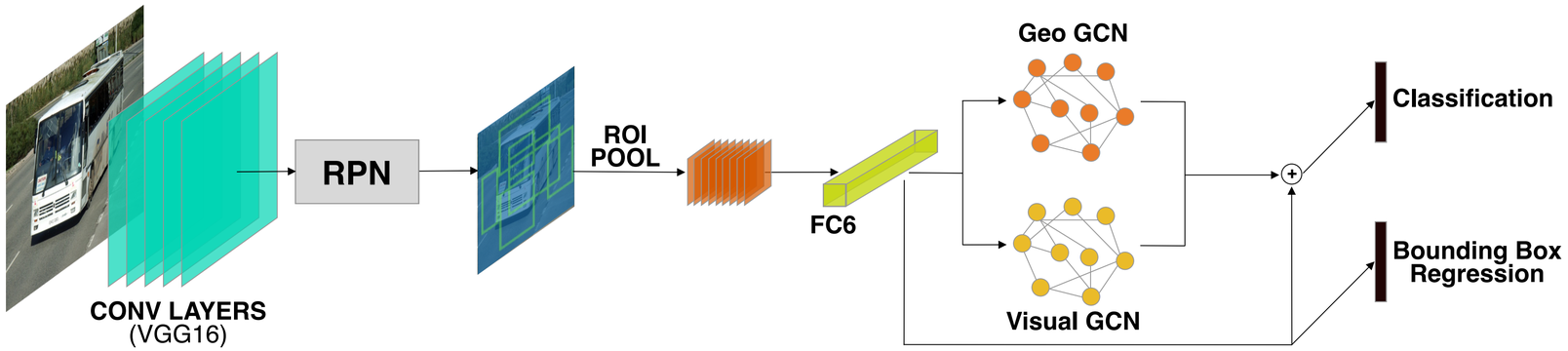
Approach

- **Base:** Models the co-occurrence of object categories based on the backbone detector's best set of guesses to capture the semantic relationship between regions.
- **Scene:** Updates scene representation at the end of each message-passing round, then uses this new representation for the next round.
- **Attr1:** Models mid-level semantic relationships between regions using object class attributes. Having built an attribute dictionary, this model learns a class-class similarity matrix over a latent attribute space and uses the highest class predictions to retrieve attribute similarity of regions.
- **Attr2:** Similar to Attr1 but first maps attributes to regions using their predicted class scores. After this mapping, it learns a region-region similarity matrix over a latent attribute space and retrieves attribute similarity of regions.



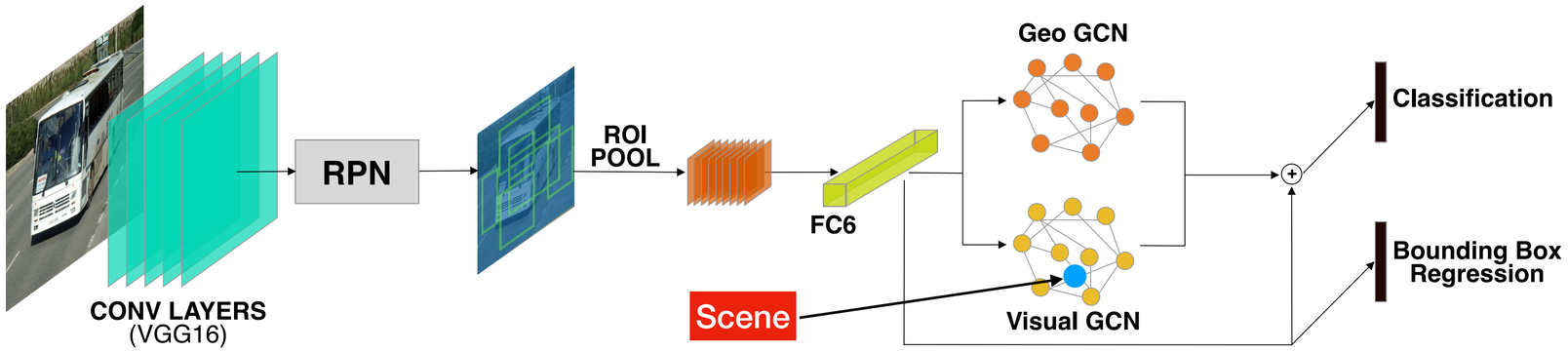
K: Number of regions

Approach



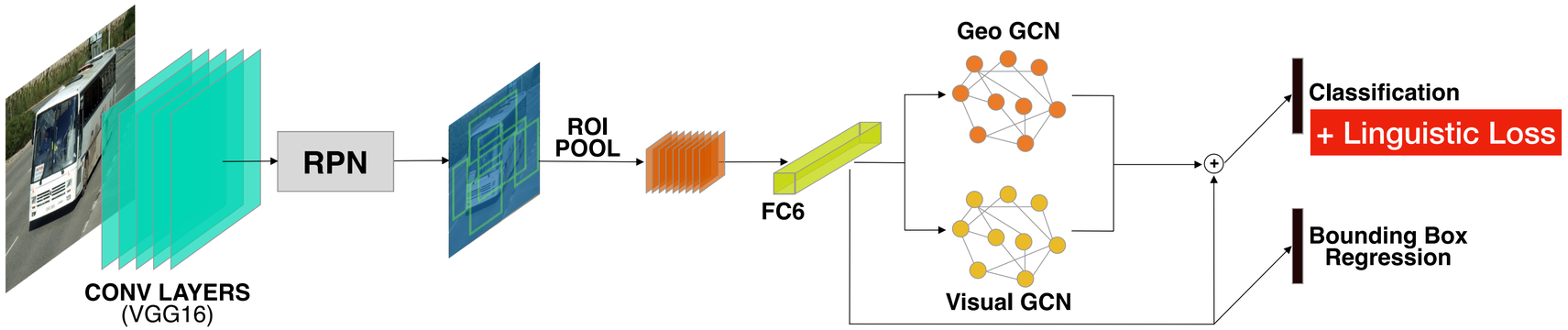
- **GeoVis:** Employs two single-layer GCNs for message-passing between regions based on their visual (Visual GCN) and spatial (Geo GCN) relationships.

Approach



- **GeoVis:** Employs two single-layer GCNs for message-passing between regions based on their visual (Visual GCN) and spatial (Geo GCN) relationships.
- **GeoVis-S:** Builds on GeoVis, and adds scene as a first-class participant in Visual GCN.

Approach

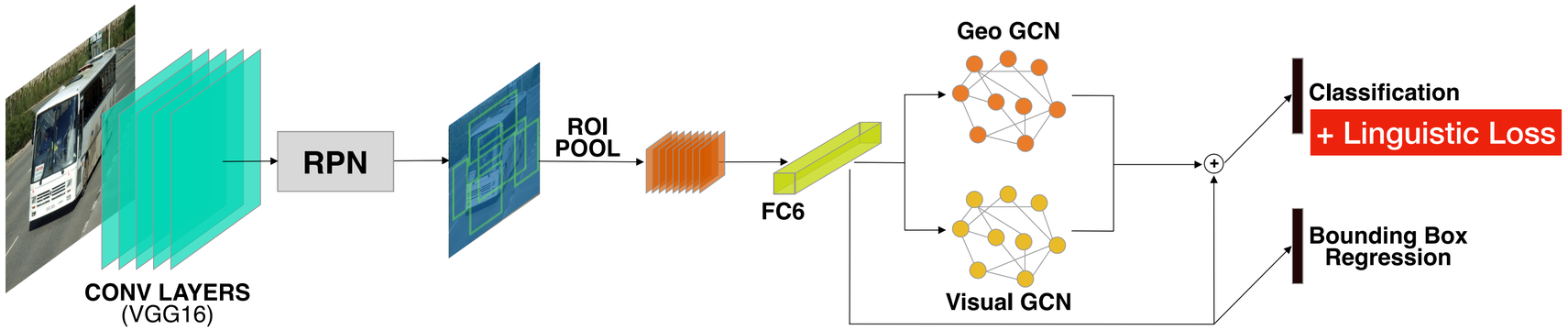


- **GeoVis:** Employs two single-layer GCNs for message-passing between regions based on their visual (Visual GCN) and spatial (Geo GCN) relationships.
- **GeoVis-S:** Builds on GeoVis, and adds scene as a first-class participant in Visual GCN.
- **GeoVis-Ling:** Uses a weighted loss formulation which penalizes misclassification of semantically similar categories more than dissimilar ones. Semantic similarity between classes is measured on a word embedding space.

Prediction
Bus

High Penalty

Approach



- **GeoVis:** Employs two single-layer GCNs for message-passing between regions based on their visual (Visual GCN) and spatial (Geo GCN) relationships.
- **GeoVis-S:** Builds on GeoVis, and adds scene as a first-class participant in Visual GCN.
- **GeoVis-Ling:** Uses a weighted loss formulation which penalizes misclassification of semantically similar categories more than dissimilar ones. Semantic similarity between classes is measured on a word embedding space.

Prediction
Dog

Low
Penalty

Experiments (Datasets)

- **PASCAL VOC**
  - **Training:** VOC 2007 trainval and VOC 2012 trainval combined
  - **Evaluation:** VOC 2007 test

- **MS COCO**
  - **Training:** COCO 2014 train
  - **Evaluation:** COCO 2014 minival and COCO 2019 test-dev (server evaluation)

Experiments (Baselines)

- **Faster R-CNN** [Ren et al., 2015]
  No explicit context modeling

- **Structure Inference Net** [Liu et al., 2018]
  Models spatial and visual relationships jointly

All models uses the same CNN backbone, VGG16, pre-trained on ImageNet.

All models were trained for the same number of steps, with the same initial learning rate, and the learning rate decaying strategy.

Results (PASCAL)

- All models outperform Faster R-CNN

	FRCNN	SIN	Scene	Attr1	Attr2	GeoVis-S	GeoVis-Ling
aeroplane	0.767	**0.780**	**0.771**	0.770	0.766	0.760	0.767
bicycle	0.793	**0.798**	**0.796**	0.789	0.789	0.795	0.788
bird	0.733	**0.765**	0.731	0.742	0.750	0.745	**0.757**
boat	0.660	**0.676**	0.668	**0.670**	**0.670**	0.629	0.639
bottle	0.611	**0.625**	0.596	0.600	0.592	**0.613**	0.608
bus	**0.853**	0.851	0.849	**0.852**	0.843	0.849	0.846
car	0.865	0.865	**0.868**	**0.866**	0.856	0.861	0.861
cat	0.881	0.870	0.882	**0.885**	**0.883**	0.881	0.876
chair	0.580	**0.615**	0.574	0.584	0.565	**0.588**	0.575
cow	0.831	0.838	0.837	0.819	0.841	**0.870**	**0.852**
diningtable	0.660	0.691	**0.721**	0.694	0.693	0.677	**0.708**
dog	0.848	0.846	**0.854**	0.852	0.850	**0.854**	0.846
horse	0.859	0.862	0.863	**0.877**	**0.871**	0.866	0.812
motorbike	0.774	**0.788**	**0.776**	0.768	0.768	0.755	0.758
person	0.782	0.786	0.785	**0.787**	**0.787**	0.781	0.778
pottedplant	0.418	**0.509**	0.443	0.429	0.446	**0.482**	0.442
sheep	0.756	**0.771**	0.769	0.752	0.756	0.760	**0.781**
sofa	0.700	**0.756**	0.732	**0.734**	0.732	0.720	0.722
train	0.821	**0.842**	0.839	0.818	**0.842**	0.821	0.825
tvmonitor	0.746	**0.768**	0.762	**0.766**	0.762	0.765	0.758
average	0.747	**0.765**	**0.756**	0.753	0.753	0.754	0.750
animals	0.818	**0.825**	0.823	0.821	**0.825**	**0.829**	0.821

Results (PASCAL)

- All models outperform Faster R-CNN
- Modeling scene as a first-class participant improves the overall performance (GeoVis mAP: %74.9)

	FRCNN	SIN	Scene	Attr1	Attr2	GeoVis-S	GeoVis-Ling
aeroplane	0.767	**0.780**	**0.771**	0.770	0.766	0.760	0.767
bicycle	0.793	**0.798**	**0.796**	0.789	0.789	0.795	0.788
bird	0.733	**0.765**	0.731	0.742	0.750	0.745	**0.757**
boat	0.660	**0.676**	0.668	**0.670**	**0.670**	0.629	0.639
bottle	0.611	**0.625**	0.596	0.600	0.592	**0.613**	0.608
bus	**0.853**	0.851	0.849	**0.852**	0.843	0.849	0.846
car	0.865	0.865	**0.868**	**0.866**	0.856	0.861	0.861
cat	0.881	0.870	0.882	**0.885**	**0.883**	0.881	0.876
chair	0.580	**0.615**	0.574	0.584	0.565	**0.588**	0.575
cow	0.831	0.838	0.837	0.819	0.841	**0.870**	**0.852**
diningtable	0.660	0.691	**0.721**	0.694	0.693	0.677	**0.708**
dog	0.848	0.846	**0.854**	0.852	0.850	**0.854**	0.846
horse	0.859	0.862	0.863	**0.877**	**0.871**	0.866	0.812
motorbike	0.774	**0.788**	**0.776**	0.768	0.768	0.755	0.758
person	0.782	0.786	0.785	**0.787**	**0.787**	0.781	0.778
pottedplant	0.418	**0.509**	0.443	0.429	0.446	**0.482**	0.442
sheep	0.756	**0.771**	0.769	0.752	0.756	0.760	**0.781**
sofa	0.700	**0.756**	0.732	**0.734**	0.732	0.720	0.722
train	0.821	**0.842**	0.839	0.818	**0.842**	0.821	0.825
tvmonitor	0.746	**0.768**	0.762	**0.766**	0.762	0.765	0.758
average	0.747	**0.765**	**0.756**	0.753	0.753	0.754	0.750
animals	0.818	**0.825**	0.823	0.821	**0.825**	**0.829**	0.821

Results (PASCAL)

- All models outperform Faster R-CNN
- Modeling scene as a first-class participant improves the overall performance (GeoVis mAP: %74.9)
- Utilizing mid-level semantic cues and semantic-aware loss works better for some categories but does not improve overall performance

	FRCNN	SIN	Scene	Attr1	Attr2	GeoVis-S	GeoVis-Ling
aeroplane	0.767	**0.780**	**0.771**	0.770	0.766	0.760	0.767
bicycle	0.793	**0.798**	**0.796**	0.789	0.789	0.795	0.788
bird	0.733	**0.765**	0.731	0.742	0.750	0.745	**0.757**
boat	0.660	**0.676**	0.668	**0.670**	**0.670**	0.629	0.639
bottle	0.611	**0.625**	0.596	0.600	0.592	**0.613**	0.608
bus	**0.853**	0.851	0.849	**0.852**	0.843	0.849	0.846
car	0.865	0.865	**0.868**	**0.866**	0.856	0.861	0.861
cat	0.881	0.870	0.882	**0.885**	**0.883**	0.881	0.876
chair	0.580	**0.615**	0.574	0.584	0.565	**0.588**	0.575
cow	0.831	0.838	0.837	0.819	0.841	**0.870**	**0.852**
diningtable	0.660	0.691	**0.721**	0.694	0.693	0.677	**0.708**
dog	0.848	0.846	**0.854**	0.852	0.850	**0.854**	0.846
horse	0.859	0.862	0.863	**0.877**	**0.871**	0.866	0.812
motorbike	0.774	**0.788**	**0.776**	0.768	0.768	0.755	0.758
person	0.782	0.786	0.785	**0.787**	**0.787**	0.781	0.778
pottedplant	0.418	**0.509**	0.443	0.429	0.446	**0.482**	0.442
sheep	0.756	**0.771**	0.769	0.752	0.756	0.760	**0.781**
sofa	0.700	**0.756**	0.732	**0.734**	0.732	0.720	0.722
train	0.821	**0.842**	0.839	0.818	**0.842**	0.821	0.825
tvmonitor	0.746	**0.768**	0.762	**0.766**	0.762	0.765	0.758
average	0.747	**0.765**	**0.756**	0.753	0.753	0.754	0.750
animals	0.818	**0.825**	0.823	0.821	**0.825**	**0.829**	0.821

Results (PASCAL)

- All models outperform Faster R-CNN
- Modeling scene as a first-class participant improves the overall performance (GeoVis mAP: %74.9)
- Utilizing mid-level semantic cues and semantic-aware loss works better for some categories but does not improve overall performance
- GeoVis-S achieves the best performance on animal category

	FRCNN	SIN	Scene	Attr1	Attr2	GeoVis-S	GeoVis-Ling
aeroplane	0.767	**0.780**	**0.771**	0.770	0.766	0.760	0.767
bicycle	0.793	**0.798**	**0.796**	0.789	0.789	0.795	0.788
bird	0.733	**0.765**	0.731	0.742	0.750	0.745	**0.757**
boat	0.660	**0.676**	0.668	**0.670**	**0.670**	0.629	0.639
bottle	0.611	**0.625**	0.596	0.600	0.592	**0.613**	0.608
bus	**0.853**	0.851	0.849	**0.852**	0.843	0.849	0.846
car	0.865	0.865	**0.868**	**0.866**	0.856	0.861	0.861
cat	0.881	0.870	0.882	**0.885**	**0.883**	0.881	0.876
chair	0.580	**0.615**	0.574	0.584	0.565	**0.588**	0.575
cow	0.831	0.838	0.837	0.819	0.841	**0.870**	**0.852**
diningtable	0.660	0.691	**0.721**	0.694	0.693	0.677	**0.708**
dog	0.848	0.846	**0.854**	0.852	0.850	**0.854**	0.846
horse	0.859	0.862	0.863	**0.877**	**0.871**	0.866	0.812
motorbike	0.774	**0.788**	**0.776**	0.768	0.768	0.755	0.758
person	0.782	0.786	0.785	**0.787**	**0.787**	0.781	0.778
pottedplant	0.418	**0.509**	0.443	0.429	0.446	**0.482**	0.442
sheep	0.756	**0.771**	0.769	0.752	0.756	0.760	**0.781**
sofa	0.700	**0.756**	0.732	**0.734**	0.732	0.720	0.722
train	0.821	**0.842**	0.839	0.818	**0.842**	0.821	0.825
tvmonitor	0.746	**0.768**	0.762	**0.766**	0.762	0.765	0.758
average	0.747	**0.765**	**0.756**	0.753	0.753	0.754	0.750
animals	0.818	**0.825**	0.823	0.821	**0.825**	**0.829**	0.821

Results (COCO)

- On COCO 2014 minival, GeoVis-S is the best performing model on 4 of the 11 supercategories while Faster R-CNN being for only 1 supercategory

COCO 2014 minival			
Test setting / Method	FRCNN	SIN	GeoVis-S
AP @[IoU=0.50:0.95 — area= all]	0.207	0.213	0.209
AP @[IoU=0.50 — area= all]	0.401	0.415	0.408
AP @[IoU=0.75 — area= all]	0.196	0.197	0.193
AP @[IoU=0.50:0.95 — area= small]	0.050	0.055	0.051
AP @[IoU=0.50:0.95 — area=medium]	0.232	0.242	0.237
AP @[IoU=0.50:0.95 — area= large]	0.342	0.346	0.345
Accessories AP @ IoU=[0.50,0.95]	0.079	0.084	0.084
Food AP @ IoU=[0.50,0.95]	0.161	0.166	0.169
Kitchenware AP @ IoU=[0.50,0.95]	0.116	0.118	0.120
Furniture AP @ IoU=[0.50,0.95]	0.216	0.225	0.216
Electronics AP @ IoU=[0.50,0.95]	0.245	0.263	0.255
Appliance AP @ IoU=[0.50,0.95]	0.228	0.252	0.215
Indoor objects AP @ IoU=[0.50,0.95]	0.133	0.137	0.138
Animal AP @ IoU=[0.50,0.95]	0.374	0.371	0.373
Vehicle AP @ IoU=[0.50,0.95]	0.262	0.265	0.262
Sports AP @ IoU=[0.50,0.95]	0.145	0.149	0.146
Outdoor objects AP @ IoU=[0.50,0.95]	0.270	0.271	0.261

COCO 2019 test-dev			
Test setting / Method	FRCNN	SIN	GeoVis-S
AP @[IoU=0.50:0.95 — area= all]	0.207	0.215	0.211
AP @[IoU=0.50 — area= all]	0.403	0.423	0.411
AP @[IoU=0.75 — area= all]	0.194	0.198	0.198

Results (COCO)

- On COCO 2014 minival, GeoVis-S is the best performing model on 4 of the 11 supercategories while Faster R-CNN being for only 1 supercategory

- On COCO 2019 test-dev, **GeoVis-S achieves the same performance as more costly SIN**, when required IoU threshold is 0.75

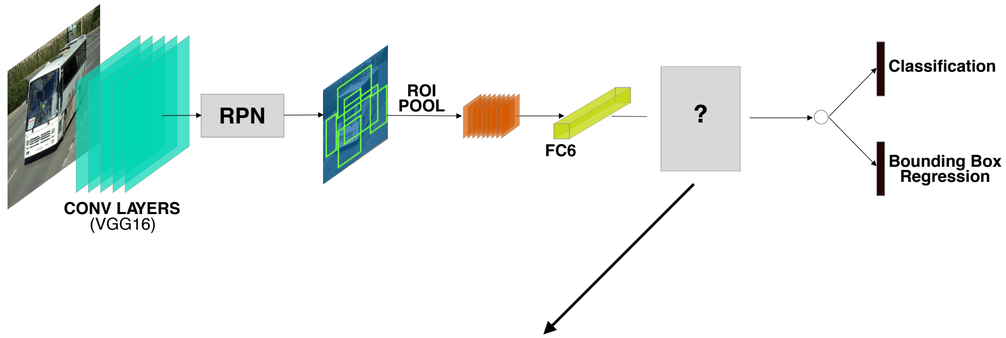
- This result indicates that our model is good at localization but may be suffering from poor classification / region proposal

COCO 2014 minival			
Test setting / Method	FRCNN	SIN	GeoVis-S
AP @[IoU=0.50:0.95 — area= all]	0.207	0.213	0.209
AP @[IoU=0.50 — area= all]	0.401	0.415	0.408
AP @[IoU=0.75 — area= all]	0.196	0.197	0.193
AP @[IoU=0.50:0.95 — area= small]	0.050	0.055	0.051
AP @[IoU=0.50:0.95 — area=medium]	0.232	0.242	0.237
AP @[IoU=0.50:0.95 — area= large]	0.342	0.346	0.345
Accessories AP @ IoU=[0.50,0.95]	0.079	0.084	0.084
Food AP @ IoU=[0.50,0.95]	0.161	0.166	0.169
Kitchenware AP @ IoU=[0.50,0.95]	0.116	0.118	0.120
Furniture AP @ IoU=[0.50,0.95]	0.216	0.225	0.216
Electronics AP @ IoU=[0.50,0.95]	0.245	0.263	0.255
Appliance AP @ IoU=[0.50,0.95]	0.228	0.252	0.215
Indoor objects AP @ IoU=[0.50,0.95]	0.133	0.137	0.138
Animal AP @ IoU=[0.50,0.95]	0.374	0.371	0.373
Vehicle AP @ IoU=[0.50,0.95]	0.262	0.265	0.262
Sports AP @ IoU=[0.50,0.95]	0.145	0.149	0.146
Outdoor objects AP @ IoU=[0.50,0.95]	0.270	0.271	0.261

COCO 2019 test-dev			
Test setting / Method	FRCNN	SIN	GeoVis-S
AP @[IoU=0.50:0.95 — area= all]	0.207	0.215	0.211
AP @[IoU=0.50 — area= all]	0.403	0.423	0.411
AP @[IoU=0.75 — area= all]	0.194	0.198	0.198

Comparison of Model Parameters

- All three models are identical up to FC6, and their R-CNN heads operate on $\mathbb{R}^{4096}$ so we can compare the number of parameters in between for a fair and dataset-agnostic comparison

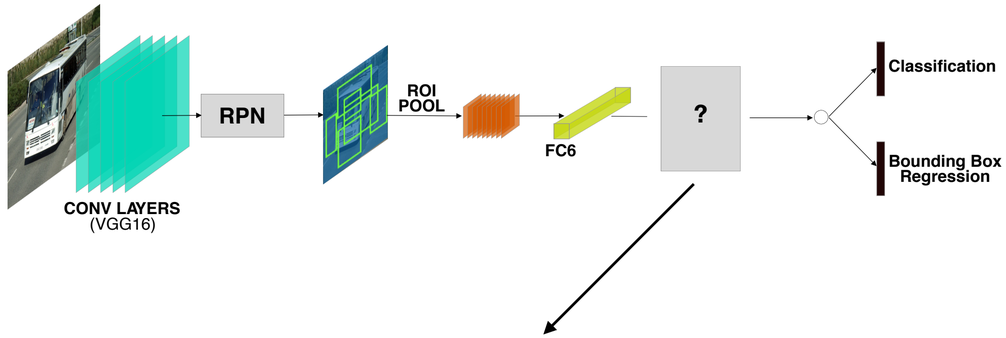


Faster R-CNN: A fully-connected layer (FC7)
SIN: Graph structure inference module
GeoVis-S: VisualGCN and GeoGCN

Comparison of Model Parameters

- All three models are identical up to FC6, and their R-CNN heads operate on $\mathbb{R}^{4096}$ so we can compare the number of parameters in between for a fair and dataset-agnostic comparison

- SIN uses almost **12x** more parameters than Faster R-CNN for context modeling

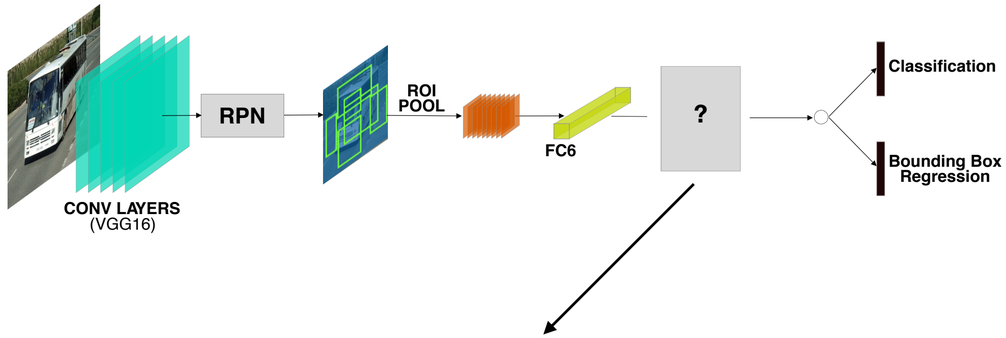


Faster R-CNN: A fully-connected layer (FC7)
SIN: Graph structure inference module
GeoVis-S: VisualGCN and GeoGCN

Method	FRCNN	SIN	GeoVis-S (Ours)
# Params	16,781,312	201,359,372	33,570,829

Comparison of Model Parameters

- All three models are identical up to FC6, and their R-CNN heads operate on $\mathbb{R}^{4096}$ so we can compare the number of parameters in between for a fair and dataset-agnostic comparison
- SIN uses almost **12x** more parameters than Faster R-CNN for context modeling
- SIN uses almost **6x** more parameters than GeoVis-S for context modeling
- Our model is more feasible to deploy on resource-constrained devices and more suitable for parallel training as it will require less bandwidth



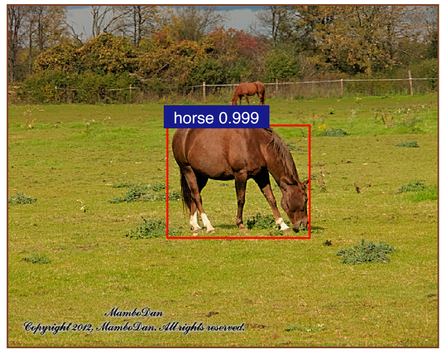
Faster R-CNN: A fully-connected layer (FC7)
SIN: Graph structure inference module
GeoVis-S: VisualGCN and GeoGCN

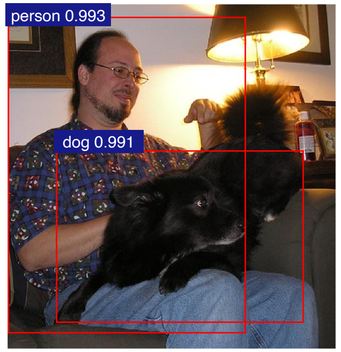
Method	FRCNN	SIN	GeoVis-S (Ours)
# Params	16,781,312	201,359,372	33,570,829

Qualitative Results

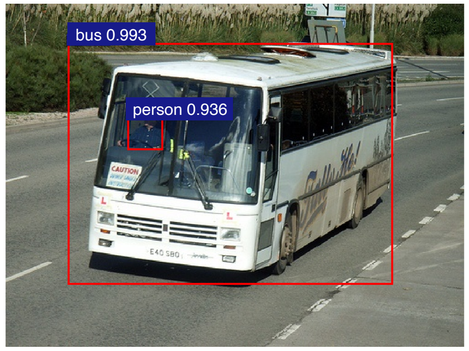
- As SIN passes messages between regions based on a single graphical representation wherein edges encode joint spatio-visual relationships between regions, it fails in utilizing context for rare object placements.

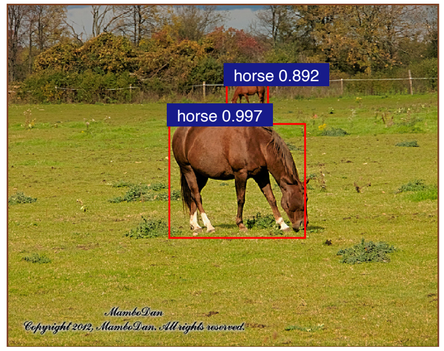
Structure Inference Net

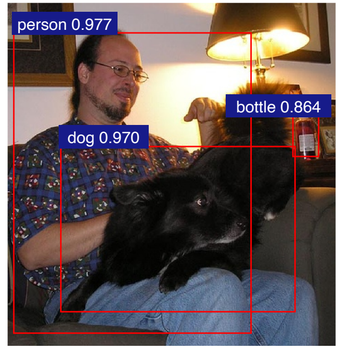


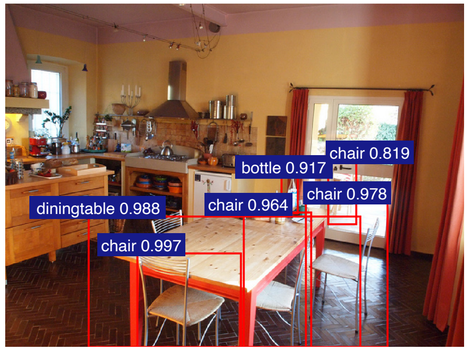


GeoVis-S (Ours)









Conclusion and Future Work

- We propose a lightweight belief-propagation mechanism for context-aware object detection. We also experiment with several semantic cues from different levels as the source of context.

- Our proposed approach adds negligible amount of extra parameters on Faster R-CNN, yet brings significant improvement in performance. It also performs competitively against more costly SIN.

- We will apply our findings on weakly-supervised detection and video detection settings.

Thank you for your attention!