Detecting and Recognizing Humans, Objects, and their Interactions

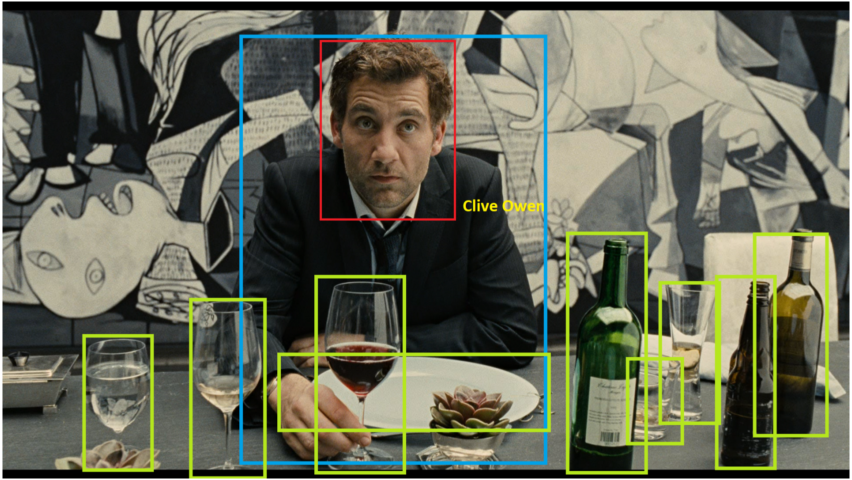
Ankan Bansal

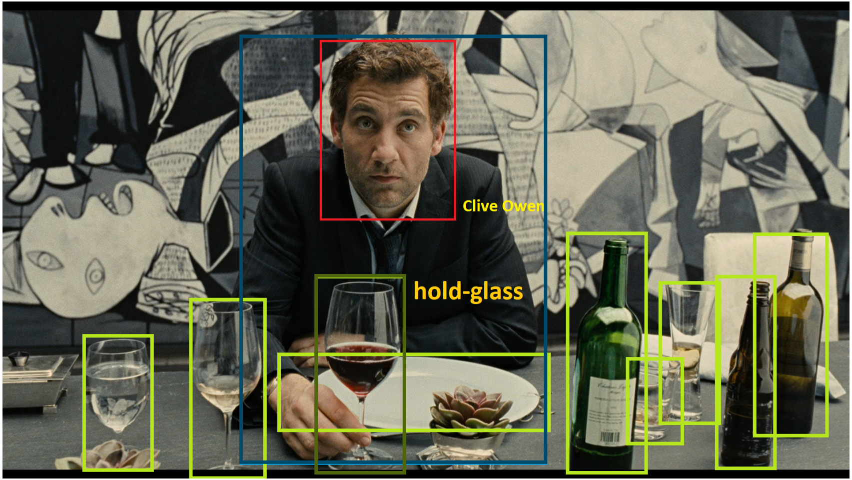
PhD Dissetation Defense

Committee:
Prof. Rama Chellappa
Prof. Ramani Duraiswami
Prof. Behtash Babadi
Prof. Abhinav Shrivastava
Prof. David Jacobs

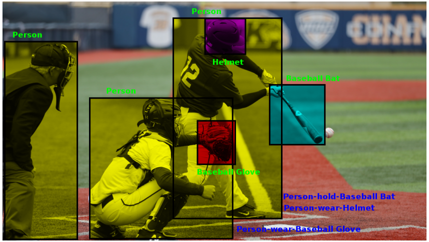
07 May, 2020


Clive Owen


Clive Owen

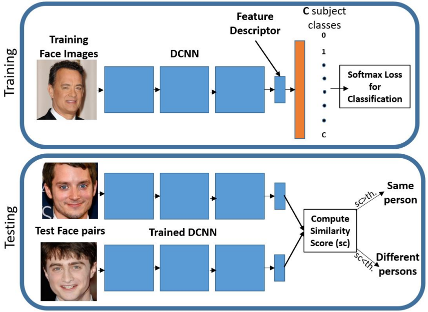

Clive Owen
hold-glass

Detecting Objects and Interactions



- Zero-Shot Object Detection. *Bansal, Sikka, Sharma, Chellappa, Divakaran*. European Conference on Computer Vision (ECCV), 2018.
- Detecting Human-Object Interactions via Functional Generalization. *Bansal, Rambhatla, Shrivastava, Chellappa*. Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI), 2020.
- Spatial Priming for Detecting Human-Object Interactions. *Bansal, Rambhatla, Shrivastava, Chellappa*. Under Submission, 2020.
- Visual Question Answering on Image Sets. *Bansal, Zhang, Chellappa*. Under Submission, 2020.

Face Recognition



- UMDFaces: An Annotated Face Dataset for Training Deep Networks. *Bansal, Nanduri, Castillo, Ranjan, Chellappa*. International Joint Conference on Biometrics (IJCB), 2017.
- The Do's and Don'ts for CNN-Based Face Verification. *Bansal, Castillo, Ranjan, Chellappa*. International Conference on Computer Vision (ICCV) Workshops, 2017.
- Deep Learning for Understanding Faces. *Ranjan, Sankaranarayanan, Bansal, Bodla, Chen, Patel, Castillo, Chellappa*. IEEE Signal Processing Magazine, 2017.
- Deep Features for Recognizing Disguised Faces in the Wild. *Bansal, Ranjan, Castillo, Chellappa*. Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.
- A Fast and Accurate System for Face Detection, Identification, and Verification. *Ranjan, Bansal, Zheng, Xu, Gleason, Lu, Nanduri, Chen, Castillo, Chellappa*. IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM), 2019.

Deep CNN-based Face Recognition

Ankan Bansal, Rajeev Ranjan, Anirudh Nanduri, Jun-Cheng Chen, Carlos Castillo, Rama Chellappa

UMDFaces

- 367,888 annotated faces
- 8,277 unique identities

UMDFaces-Videos

- 22,075 videos for 3,107 identities
- 3,735,476 annotated frames

- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

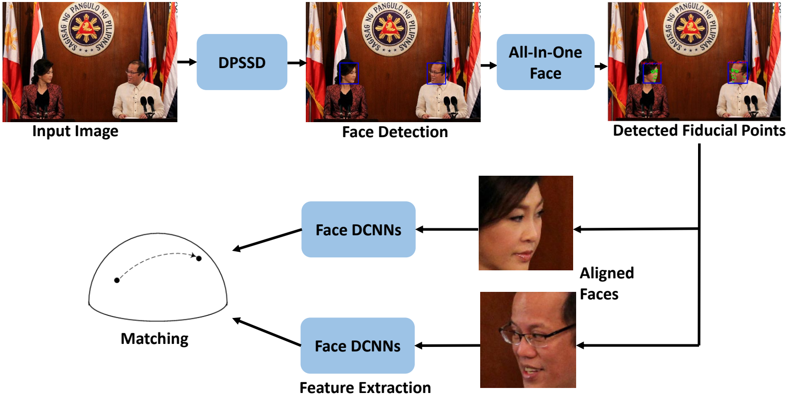
- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

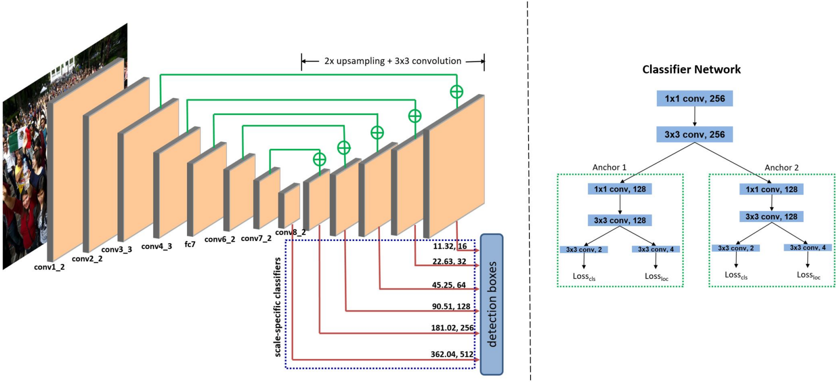
- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

- Can we train CNNs on still images and expect them to work for videos?
  No. Using mixed data is better for both mixed test datasets and video test datasets
- Are deeper datasets better than wider datasets?
  Depends on the network. Deeper datasets work well for deep networks and wide datasets work well for shallow networks
- Does label noise improve performance of deep networks?
  No. Clean data is the best
- Is alignment necessary for good performance in face recognition?
  Yes. Good keypoints and alignment lead to performance improvements

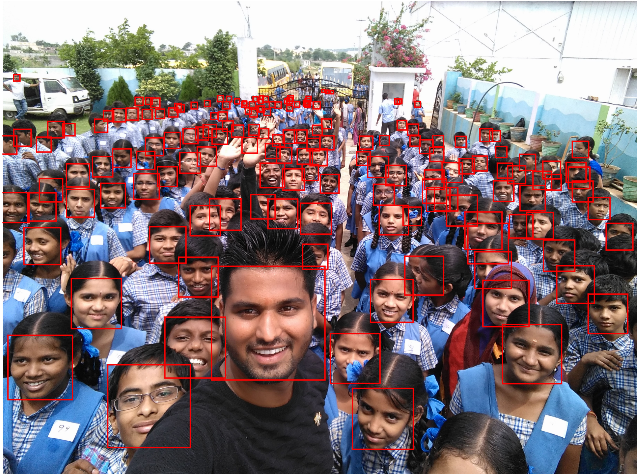
A Fast and Accurate Face Recognition System



Face Detector: DPSSD



Ranjan *et al.*, A Fast and Accurate System for Face Detection,Identification, and Verification, *T-BIOM*, 2019

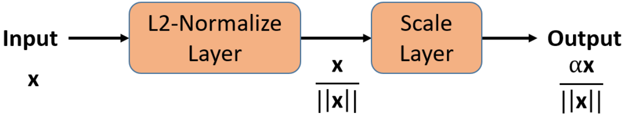


Ranjan *et al.*, A Fast and Accurate System for Face Detection,Identification, and Verification, *T-BIOM*, 2019

Crystal Loss

$$\text{minimize} \quad -\frac{1}{M}\sum_{i=1}^{M} \log \frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}}$$

$$\text{subject to} \quad \|f(\mathbf{x}_i)\|_2 = \alpha, \quad \forall i = 1, 2, \ldots M,$$



Ranjan *et al.*, A Fast and Accurate System for Face Detection,Identification, and Verification, *T-BIOM*, 2019

Inception ResNet-v2

- Input size 299×299
- Trained on Universe - MS1M + UMDFaces + UMDFaces-Videos

Evaluation

IJB-A

- 500 subjects
- 5,400 images, 2,000 videos split into 20,400 frames

IJB-B

- 1,800 subjects
- 22,000 images, 55,000 video frames
- 8,000,000 imposter pairs and 10,270 genuine pairs for 1:1 verification

	TAR (%) @ FAR			
Method	0.0001	0.001	0.01	0.1
Wang *et al.* - Casia	-	51.4	73.2	89.5
AbdAlmageed *et al.* 2016	-	-	78.7	91.1
NAN	-	88.1	94.1	97.8
Masi *at al.* 2016	-	72.5	88.6	-
Chen *et al.* DCNN_{fusion}	-	76.0	88.9	96.8
DCNN_{tpe}	-	81.3	90.0	96.4
DCNN_{all}	-	78.7	89.3	96.8
All-In-One	-	82.3	92.2	97.6
Template Adaptation	-	-	93.9	-
RX101_{l2+tpe}	90.9	94.3	**97.0**	**98.4**
Ours	**91.7**	**95.3**	96.8	98.3

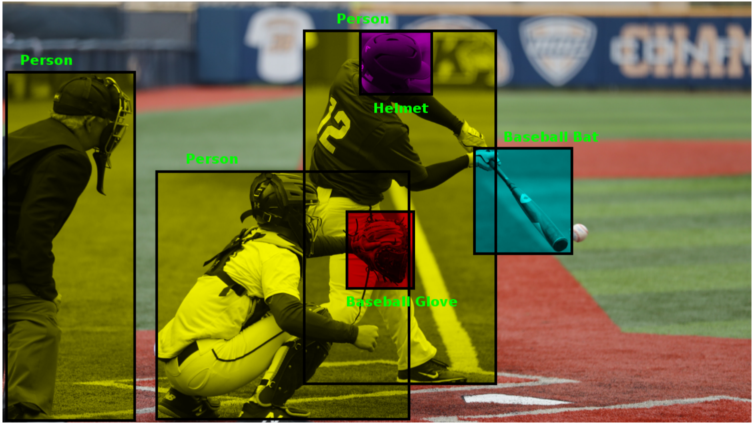
Table: IJB-A 1:1 Verification

	TAR (%) @ FAR					
Method	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
VGGFaces	-	-	55.0	72.0	86.0	-
FacePoseNet	-	-	83.2	91.6	96.5	-
Light CNN-29	-	-	87.7	92.0	95.3	-
VGGFace2	-	70.5	83.1	90.8	95.6	-
Center Loss	31.0	63.6	80.7	90.0	95.1	98.4
MN-vc	-	-	83.1	90.9	95.8	98.5
SENet50+DCN	-	-	84.9	93.7	**97.5**	**99.7**
ArcFace	37.5	**89.0**	**94.2**	**96.0**	<u>97.5</u>	98.4
Ours	27.7	61.6	89.1	94.3	97.0	98.7

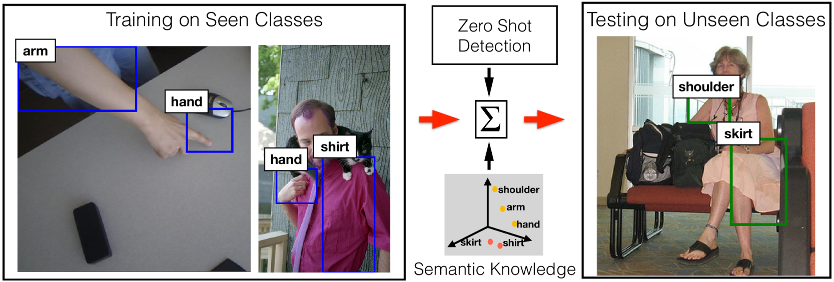
Table: IJB-B 1:1 Verification

Zero-Shot Object Detection

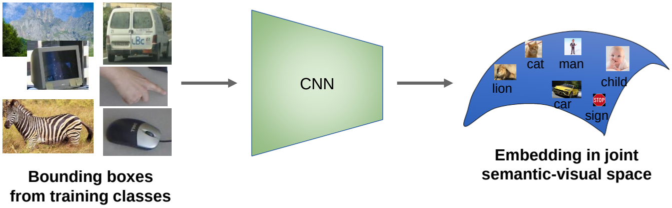
Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, Ajay Divakaran


Person
Person
Helmet
Baseball Bat
Person
Baseball Glove

- Images from few (seen) classes available for training
- Test on unseen classes



Baseline Approach



- Project bounding box, b_i, to the joint semantic-visual space $\psi_i = W_p\phi(b_i)$
- Maximize cosine similarity between ψ_i, and class embeddings, w_i

Ranking Loss

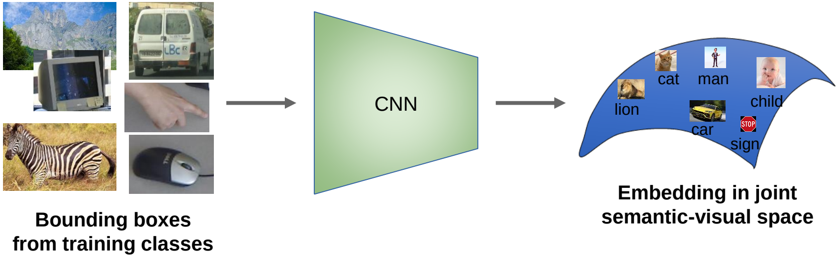
$$\mathcal{L}(b_i, y_i, \theta) = \sum_{j \in \mathcal{S}, j \neq i} \max(0, m - S_{ii} + S_{ij})$$

where $\mathcal{S}$ is the set of all seen classes

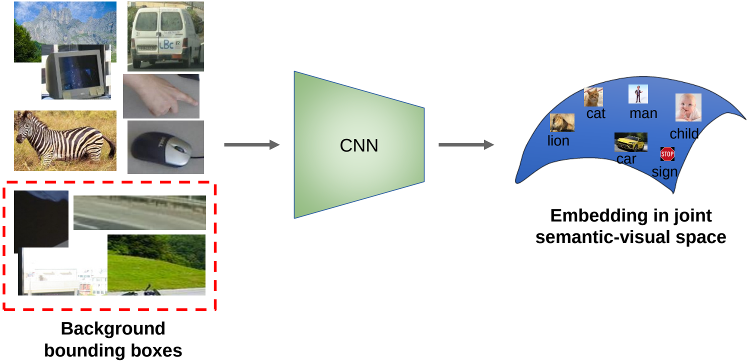
Prediction

$$\hat{y}_i = \text{argmax}_{j \in \mathcal{U}} S_{ij}$$

where $\mathcal{U}$ is the set of test classes


Bounding boxes
from training classes
CNN
cat
man
lion
child
car
sign
Embedding in joint
semantic-visual space

Statically Assigned Background (SB)



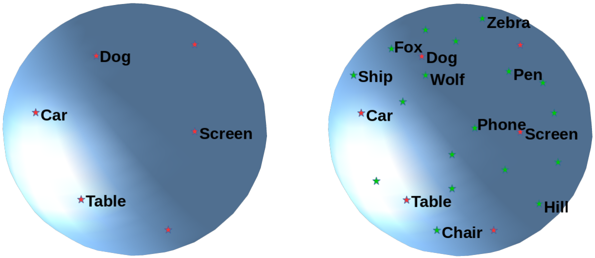
- Add a fixed "background" label and assign all background boxes to this label

Latent Assignment Based (LAB) ZSD

- Spread background boxes across the embedding space
- EM-like approach that iteratively
  - (1) assigns classes from a large vocabulary to background boxes, and
  - (2) fine-tunes the model
- Background boxes could possibly belong to the "background set"

Densely Sampled Embedding Space (DSES)

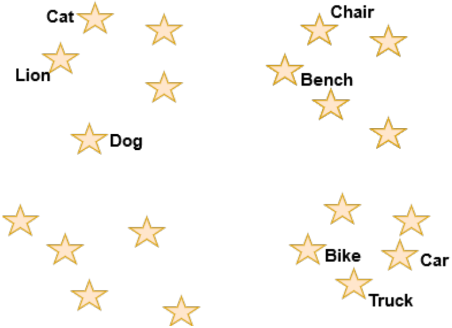
- Joint embeddings suffer from sparse sampling in the visual-semantic space
- Augment labels with OpenImages (OI)



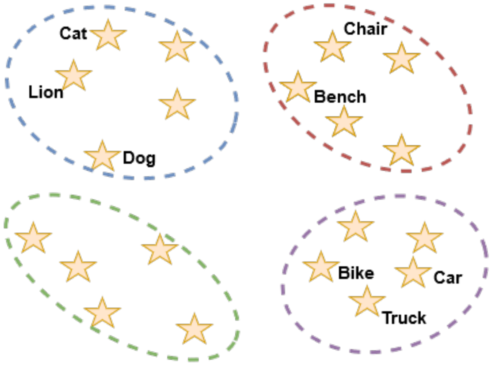
Datasets

Dataset	Seen Classes	Unseen Classes	# Training Boxes
MSCOCO	48	17	1.4 million
VisualGenome (VG)	478	130	5.8 million
OpenImages	545	-	400 thousand

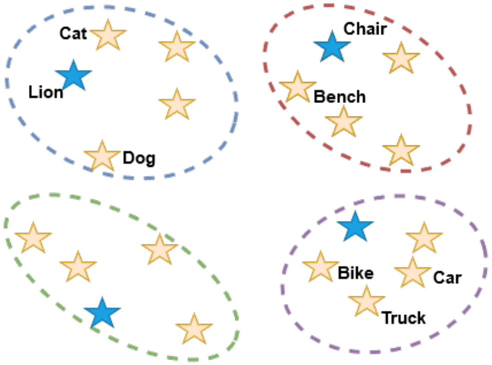
Train and Test Splits



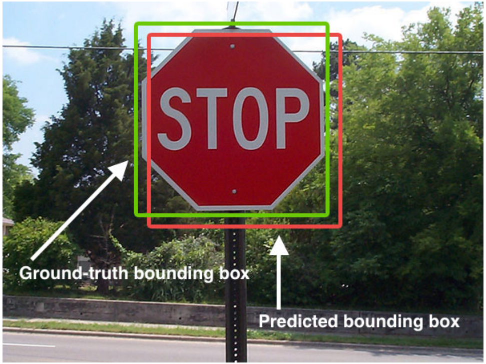
Train and Test Splits



Train and Test Splits



Evaluation - IoU



$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Evaluation

- Recall@K $\rightarrow$ Recall when the top K boxes are selected
- $K = 100$

ZSD method	BG-aware	MSCOCO #classes $\|\mathcal{S}\|$	$\|\mathcal{U}\|$	$\|\mathcal{O}\|$	MSCOCO IoU 0.4	0.5	0.6	VisualGenome #classes $\|\mathcal{S}\|$	$\|\mathcal{U}\|$	$\|\mathcal{O}\|$	VisualGenome IoU 0.4	0.5	0.6
Baseline		48	17	0	34.36	22.14	11.31	478	130	0	8.19	5.19	2.63
SB	✓	48	17	1	34.46	24.39	12.55	478	130	1	6.06	4.09	2.43
DSES		378	17	0	**40.23**	**27.19**	**13.63**	716	130	0	7.78	4.75	2.34
LAB	✓	48	17	343	31.86	20.52	9.98	478	130	1673	**8.43**	**5.40**	**2.74**

Table: Recall@100 performance (%) for different methods at several IoU thresholds. $|\mathcal{S}|$, $|\mathcal{U}|$, and $|\mathcal{O}|$ are the number of seen, unseen and the average number of active background classes considered during training respectively.

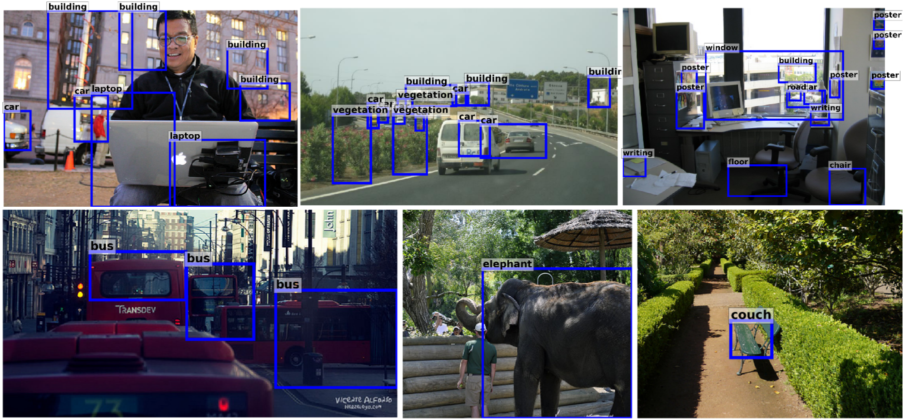
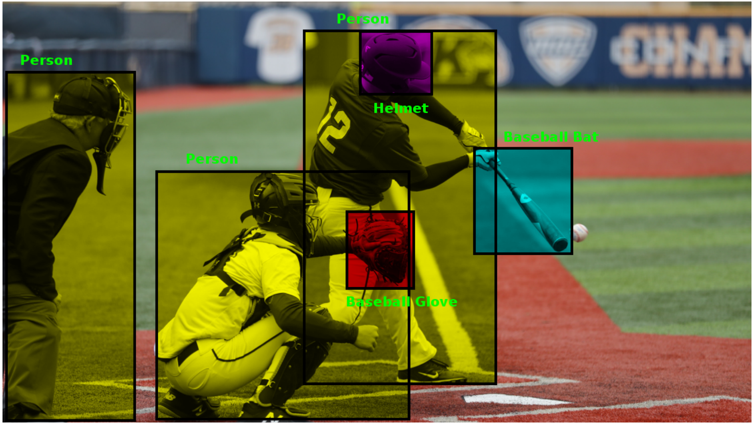
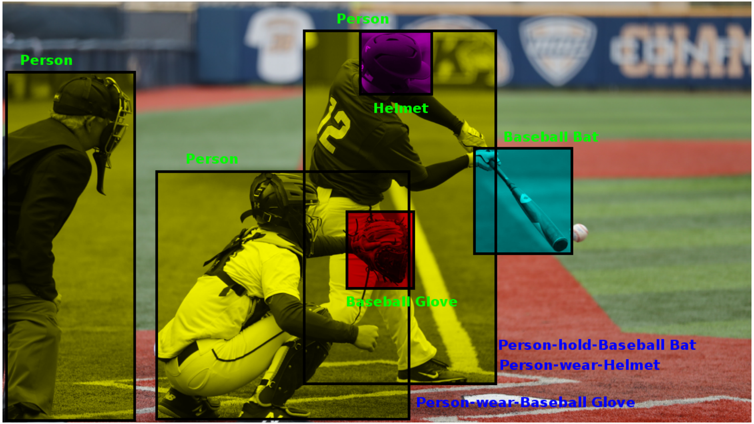


Figure: Results for LAB for VisualGenome (row 1) and SB model (row 2) for MSCOCO.

Object Detection



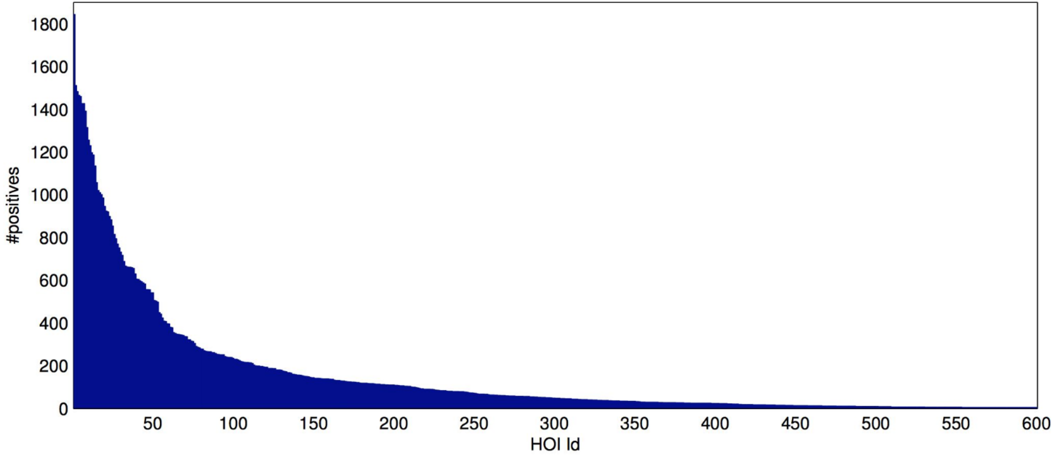
Human-Object Interaction Detection



Detecting Human-Object Interactions via Functional Generalization

Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, Rama Chellappa

Triplets of the form: $\langle \texttt{human}, \texttt{predicate}, \texttt{object} \rangle$


1800
1600
1400
1200
1000
800
600
400
200
0
#positives
50
100
150
200
250
300
350
400
450
500
550
600
HOI Id

⟨human, ride, bicycle⟩
⟨human, sit_on, bicycle⟩
⟨human, straddle, bicycle⟩

⟨human, sit_on, bicycle⟩

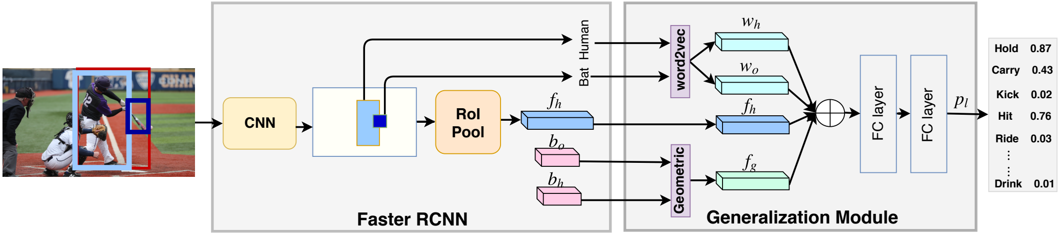

Coca-Cola
cherry

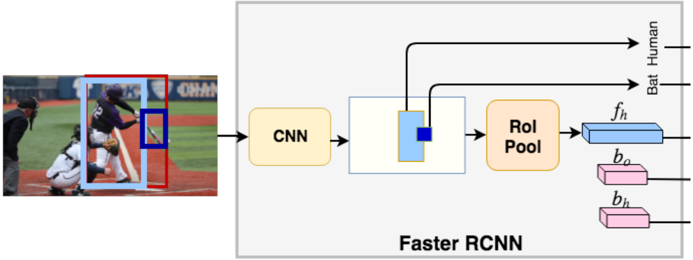

can?
mug?
cup?
glass?

Humans interaction with functionally similar objects in a similar manner


horse?
camel?
elephant?
zebra?
giraffe?
sheep?
goat?
deer?
bull?
horse?
zebra?
elephant?
can?
glass?
cup?
mug?
can?
bottle?
glass?
can?
cup?
mug?
glass?
pear?
peach?
potato?
fig?
bell pepper?
cucumber?
potato?
mango?
apple?
pear?
peach?
fig?
apple?
carrot?
cake?
muffin?
donut?
muffin?
cake?
cake?
muffin?
cheese?
donut?

- Humans interact with similar objects similarly
- Additional data obtained by replacing objects by functionally similar objects

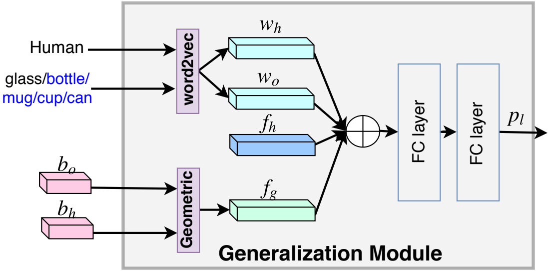

CNN
RoI
Pool
Human
Bat
f_h
b_o
b_h
Faster RCNN
word2vec
w_h
w_o
f_h
Geometric
f_g
FC layer
FC layer
p_l
Generalization Module
Hold 0.87
Carry 0.43
Kick 0.02
Hit 0.76
Ride 0.03
Drink 0.01



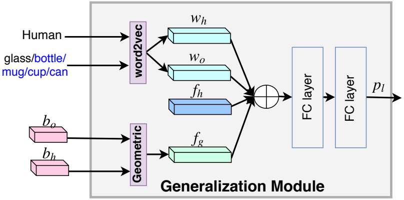
- Outputs the bounding boxes, object classes, and ROI pooled features

Functional Generalization

- An object can be replaced by a functionally similar object



Functional Generalization Module



Word Embeddings

- 300-D vectors, w_h and w_o, from `word2vec`
- Incorporate semantic information

Layout Features

- f_g encodes the relative sizes and orientations of b_h and b_o
- 14-D feature

Layout Feature

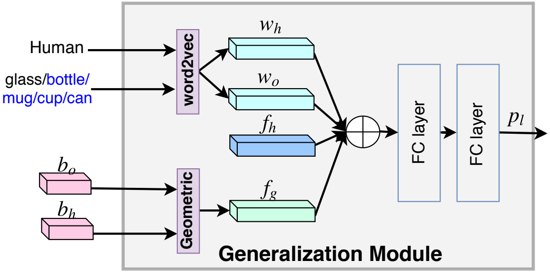
$$f_g = \left[\frac{x_1^h}{W}, \frac{y_1^h}{H}, \frac{x_2^h}{W}, \frac{y_2^h}{H}, \frac{A^h}{A^I}, \frac{x_1^o}{W}, \frac{y_1^o}{H}, \frac{x_2^o}{W}, \frac{y_2^o}{H}, \frac{A^o}{A^I}, \left(\frac{x_1^h - x_1^o}{x_2^o - x_1^o} \right), \left(\frac{y_1^h - y_1^o}{y_2^o - y_1^o} \right), \log\left(\frac{x_2^h - x_1^h}{x_2^o - x_1^o} \right), \log\left(\frac{y_2^h - y_1^h}{y_2^o - y_1^o} \right) \right]$$

Finding Similar Objects

- WordNet hierarchy?
- Large vocabulary of objects $\mathcal{V} = \{o_1, o_2, \ldots, o_n\}$
- Concatenate visual features and word vectors
- Cluster into k clusters
- Objects in the same cluster are functionally similar

- ‘Pitcher’, ‘Teapot’, ‘Kettle’, ‘Jug’
- ‘Elephant’, ‘Dinosaur’, ‘Horse’, ‘Zebra’, ‘Mule’, ‘Camel’, ‘Bull’
- ‘Can’, ‘Cup’, ‘Glass’, ‘Bottle’
- ‘Cake’, ‘Muffin’, ‘Cheese’, ‘Donut’
- ‘Apple’, ‘Pear’, ‘Peach’, ‘Fig’

Functional Generalization



- $\langle$human, drink, glass$\rangle \rightarrow \langle$human, drink, cup$\rangle$, $\langle$human, drink, can$\rangle$
- $\langle$human, ride, elephant$\rangle \rightarrow \langle$human, ride, horse$\rangle$, $\langle$human, ride, camel$\rangle$

HICO-Det

- 600 HOI triplet categories for 80 objects
- Training set - $38,000$ images with $120,000$ HOI annotations
- Test set - $9,600$ images with $33,400$ HOI instances

- Metric: Mean Average Precision (mAP %)
- Three settings: Full, Rare, Non-rare

Method	**Full** (600 classes)	**Rare** (138 classes)	**Non-rare** (462 classes)
Shen *et al.*	6.46	4.24	7.12
HO-RCNN + IP	7.30	4.68	8.08
HO-RCNN + IP + S	7.81	5.37	8.54
InteractNet	9.94	7.16	10.77
iHOI	9.97	7.11	10.83
GPNN	13.11	9.34	14.23
ICAN	14.84	10.45	16.15
Gupta *et al.*	17.18	12.17	18.68
Interactiveness Prior	17.22	13.51	18.32
Peyre *et al.*	19.40	15.40	20.75
Functional Generalization (Ours)	**21.96**	**16.43**	**23.62**

Table: mAP (%) for the HICO-Det dataset.

Zero-Shot HOI Detection

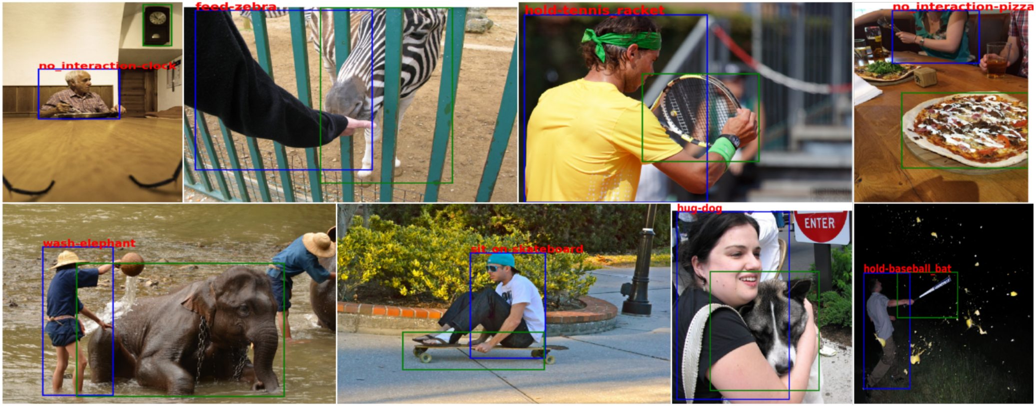
- Seen object setting
  At least one interaction seen for each object
- Unseen object setting
  No interactions seen for some object classes

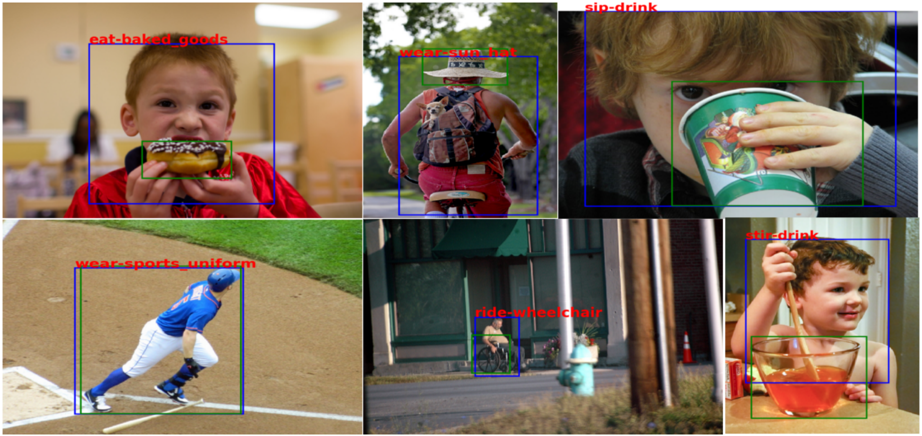
Method	**Unseen** (120 classes)	**Seen** (480 classes)	**All** (600 classes)
Shen *et al.*	5.62	-	6.26
Ours	**10.93**	12.60	**12.26**

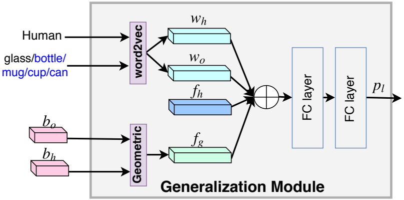
Table: Performance (mAP %) in the seen object zero-shot setting

Method	**Unseen** (100 classes)	**Seen** (500 classes)	**All** (600 classes)
Ours	11.22	14.36	13.84

Table: Performance (mAP %) in the unseen object setting


no_interaction-clock
feed-zebra
hold-tennis_racket
no_interaction-pizza
wash-elephant
sit_on-skateboard
hug-dog
ENTER
hold-baseball_bat


eat-baked_goods
wear-sun_hat
sip-drink
wear-sports_uniform
ride-wheelchair
stir-drink



Layout Features

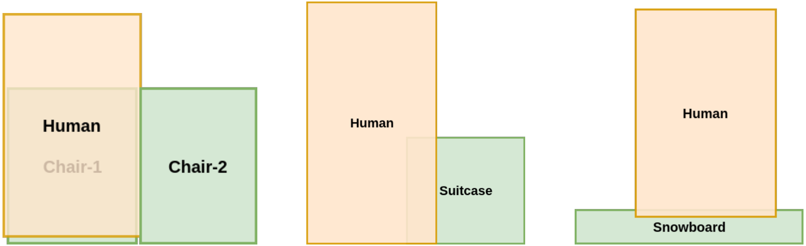
- f_g encodes the relative sizes and orientations of b_h and b_o
- 14-D feature

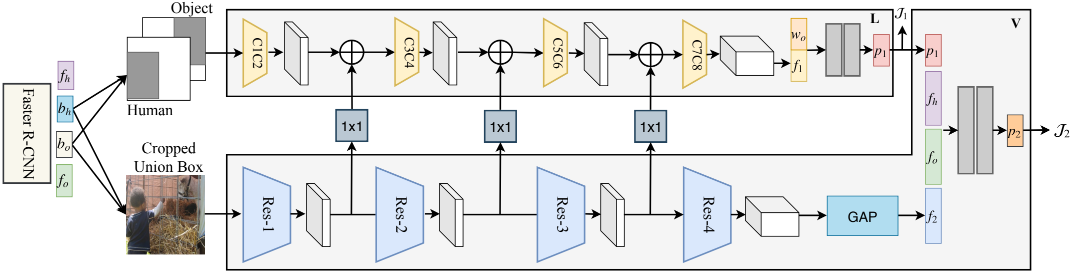
$$f_g = \left[\frac{x_1^h}{W}, \frac{y_1^h}{H}, \frac{x_2^h}{W}, \frac{y_2^h}{H}, \frac{A^h}{A^I}, \frac{x_1^o}{W}, \frac{y_1^o}{H}, \frac{x_2^o}{W}, \frac{y_2^o}{H}, \frac{A^o}{A^I}, \left(\frac{x_1^h - x_1^o}{x_2^o - x_1^o}\right), \left(\frac{y_1^h - y_1^o}{y_2^o - y_1^o}\right), \log\left(\frac{x_2^h - x_1^h}{x_2^o - x_1^o}\right), \log\left(\frac{y_2^h - y_1^h}{y_2^o - y_1^o}\right) \right]$$

Spatial Priming for Detecting Human-Object Interactions

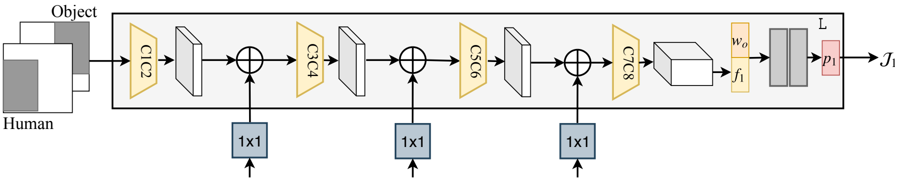
Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, Rama Chellappa

- Relative location of human and object provides useful clues
- Can make guesses based on the layout


Human
Chair-1
Chair-2
Human
Suitcase
Human
Snowboard

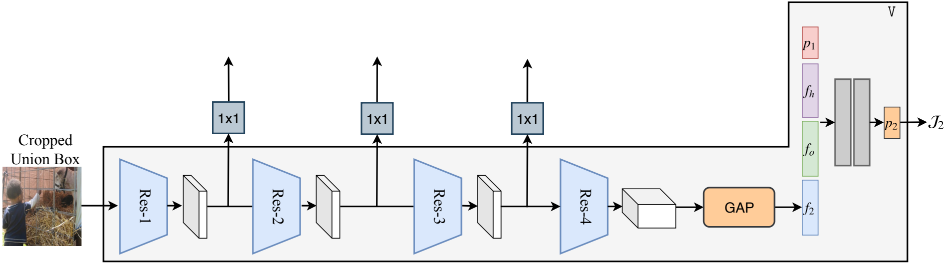

Faster R-CNN
f_h
b_h
b_o
f_o
Object
Human
Cropped Union Box
C1C2
C3C4
C5C6
C7C8
w_o
f_1
p_1
L
$\mathcal{J}_1$
1x1
1x1
1x1
Res-1
Res-2
Res-3
Res-4
GAP
f_2
V
p_1
f_h
f_o
p_2
$\mathcal{J}_2$

Layout Module



- Lateral connections from the visual module for visual context
- Semantic knowledge in the form of `word2vec` word vectors

Visual Module



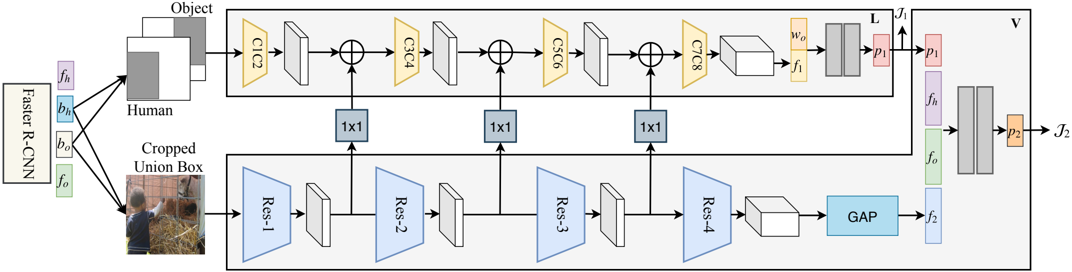
- Uses outputs of the layout module and features from the object detector

Lateral Connections

- Explicitly share visual context not available in the layout module

Spatial Priming

- Predictions from `L` prime the visual module
- Refined by the visual module


Faster R-CNN
f_h
b_h
b_o
f_o
Object
Human
Cropped
Union Box
C1C2
C3C4
C5C6
C7C8
w_o
f_1
p_1
L
$\mathcal{J}_1$
V
1x1
1x1
1x1
Res-1
Res-2
Res-3
Res-4
GAP
f_2
p_1
f_h
f_o
p_2
$\mathcal{J}_2$

HICO-Det

- 600 HOI triplet categories for 80 objects
- Training set - $38,000$ images with $120,000$ HOI annotations
- Test set - $9,600$ images with $33,400$ HOI instances

- Metric: Mean Average Precision (mAP %)
- Three settings: Full, Rare, Non-rare

Method	**Full** (600 classes)	**Rare** (138 classes)	**Non-rare** (462 classes)
Shen *et al.*	6.46	4.24	7.12
HO-RCNN + IP	7.30	4.68	8.08
HO-RCNN + IP + S	7.81	5.37	8.54
InteractNet	9.94	7.16	10.77
iHOI	9.97	7.11	10.83
GPNN	13.11	9.34	14.23
ICAN	14.84	10.45	16.15
Gupta *et al.*	17.18	12.17	18.68
Interactiveness Prior	17.22	13.51	18.32
Peyre *et al.*	19.40	15.40	20.75
Functional Generalization (Ours)	21.96	**16.43**	23.62
Spatial Priming (Ours)	**24.79**	14.77	**27.79**

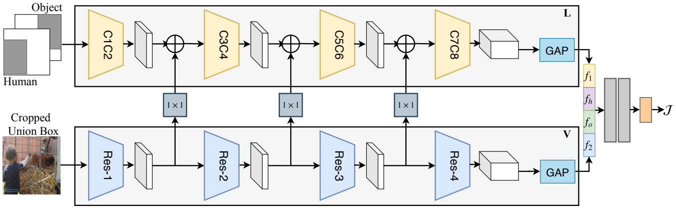
Table: Performance (mAP %) on HICO-Det

Zero-Shot HOI Detection

Method	**Unseen** (120 classes)	**Seen** (480 classes)	**All** (600 classes)
Shen *et al.*	5.62	-	6.26
Func. Gen. (Ours)	10.93	12.60	12.26
Ours	**11.06**	**21.41**	**19.34**

Table: Zero-shot HOI detection (mAP %)

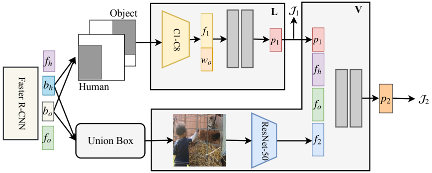
No Priming



Method	Full (600 classes)	Rare (138 classes)	Non-rare (462 classes)
`V-L`-add (NP)	**23.41**	12.14	**26.78**
NC	22.56	**12.78**	25.48
`L-V`	22.45	12.23	25.50
`V-L`-concat	22.76	11.78	26.04

Table: mAP % for the model without priming (NP). NC is same model without lateral connections

No Lateral Connections



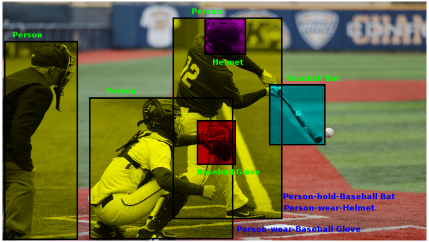
Method	Model	**Full** (600 classes)	**Rare** (138 classes)	**Non-rare** (462 classes)
NL	L	18.35	8.20	21.38
	V	**23.90**	10.82	**27.81**
NL - f_h - f_o	L	17.44	10.14	19.62
	V	23.19	**14.71**	25.72
NL - w_o	L	16.33	8.45	18.69
	V	22.91	11.29	26.39

Table: Performance (mAP %) for the model without lateral connections

Visual Question Answering on Image Sets

Ankan Bansal, Yuting Zhang, Rama Chellappa

Visual Question Answering on Image Sets



- Answer questions about a set of images
- Relate objects in one or more images
- Dataset - indoor and outdoor scenes
- VQA baselines

what the largest object in the room?

what is above the toilet wall?

what kind of car is in front of the white car?

Conclusion

- Datasets and deep networks for face recognition
- Additional sources of information for vision problems
- Semantic information from large-scale text data
- Data augmentation strategies using semantic information
- Deep encoding of geometric layout

Conclusion

- Datasets and deep networks for face recognition
- Additional sources of information for vision problems
- Semantic information from large-scale text data
- Data augmentation strategies using semantic information
- Deep encoding of geometric layout

Conclusion

- Datasets and deep networks for face recognition
- Additional sources of information for vision problems
- Semantic information from large-scale text data
- Data augmentation strategies using semantic information
- Deep encoding of geometric layout

Conclusion

- Datasets and deep networks for face recognition
- Additional sources of information for vision problems
- Semantic information from large-scale text data
- Data augmentation strategies using semantic information
- Deep encoding of geometric layout

Conclusion

- Datasets and deep networks for face recognition
- Additional sources of information for vision problems
- Semantic information from large-scale text data
- Data augmentation strategies using semantic information
- Deep encoding of geometric layout

Conclusion

- Datasets and deep networks for face recognition
- Additional sources of information for vision problems
- Semantic information from large-scale text data
- Data augmentation strategies using semantic information
- Deep encoding of geometric layout

Ideas for Future Work

- Frame semantics for video understanding
- Annotated artwork for HOI detection
- Lexical ontology and hierarchical prediction for ZSD
- Better BERT-type models for jointly learning visual-semantic spaces

Ideas for Future Work

- Frame semantics for video understanding
- Annotated artwork for HOI detection
- Lexical ontology and hierarchical prediction for ZSD
- Better BERT-type models for jointly learning visual-semantic spaces

Ideas for Future Work

- Frame semantics for video understanding
- Annotated artwork for HOI detection
- Lexical ontology and hierarchical prediction for ZSD
- Better BERT-type models for jointly learning visual-semantic spaces

Ideas for Future Work

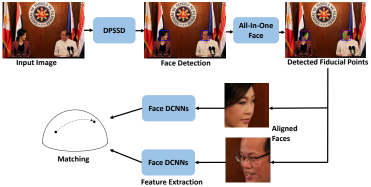
- Frame semantics for video understanding
- Annotated artwork for HOI detection
- Lexical ontology and hierarchical prediction for ZSD
- Better BERT-type models for jointly learning visual-semantic spaces

Ideas for Future Work

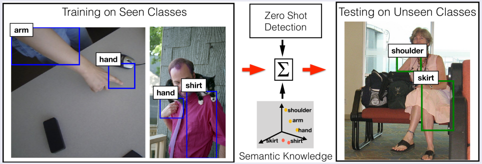
- Frame semantics for video understanding
- Annotated artwork for HOI detection
- Lexical ontology and hierarchical prediction for ZSD
- Better BERT-type models for jointly learning visual-semantic spaces

Questions?

Face Recognition



Zero-Shot Object Detection



Functional Generalization

Spatial Priming for HOI Detection