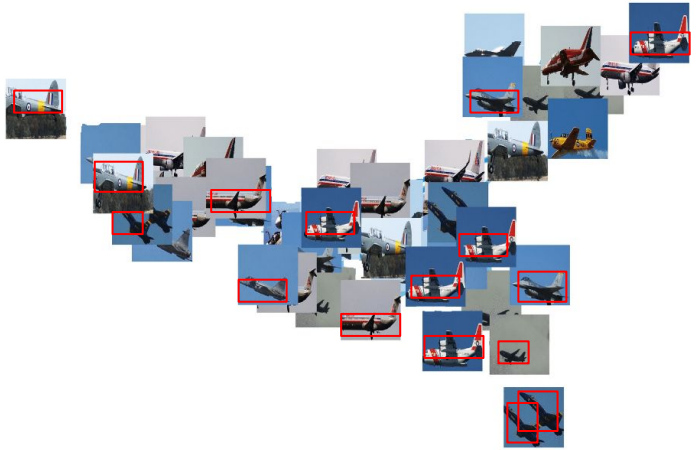
Large-Scale Unsupervised Object Discovery Vo et al. 2021

Seminar in Computer Vision, University of Freiburg, SS22

Uns

- 

- prominent objects are objects occurring in several images

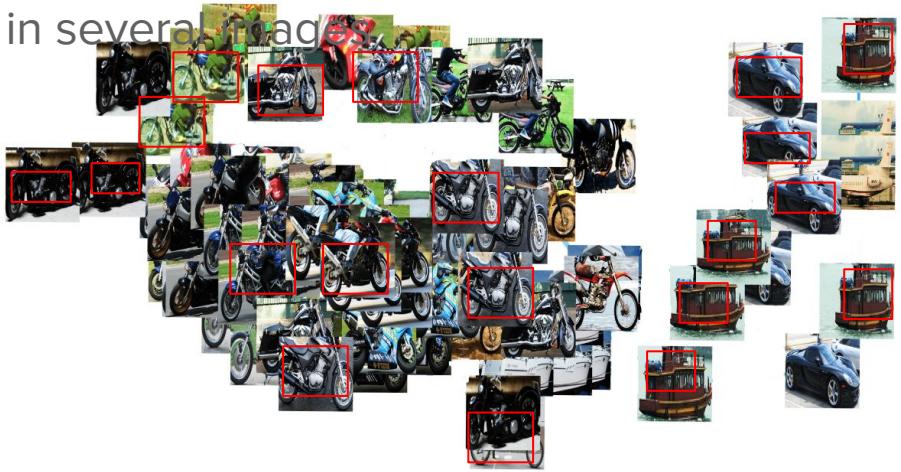


Image source [3]

Unsupervised Object Discovery

- Can typically use Object Detection datasets
- Ground truth bounding boxes are considered true objects
- Ground truth boxes used only for evaluation

Image source [3]

Outline

- **Related works**
- Methodology
- Baselines
- Experimental Settings
- Results & Conclusion

Related Works

- Early approaches → Small-scale datasets
  - Probabilistic Models
  - Non-negative matrix factorization (NMF)
  - Clustering

- Recent approaches → Moderate-sized datasets
  - Assume images are embed in a graph structure
  - Candidate region proposals generated for each image
  - Combinatorial optimization to select good proposals

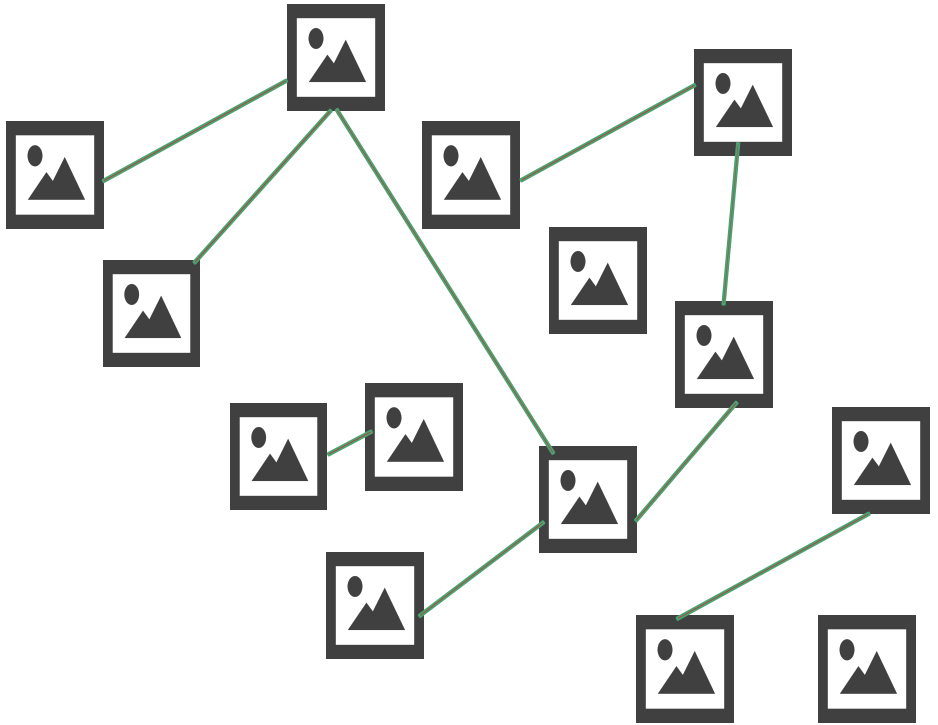
- Other approaches → Small-scale datasets
  - Decompose image into objects
  - Learn image representation

Outline

- Related works
- **Methodology**
- Baselines
- Experimental Settings
- Results & Conclusion

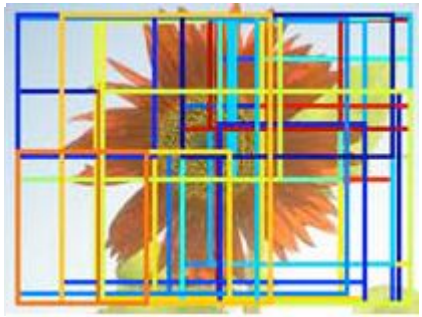
Methodology: Previous work (Vo et al. 2019)

- Initial work that scales up to moderate-size datasets
- Assume Implicit graph structure over Images
- Edges connect visually similar images

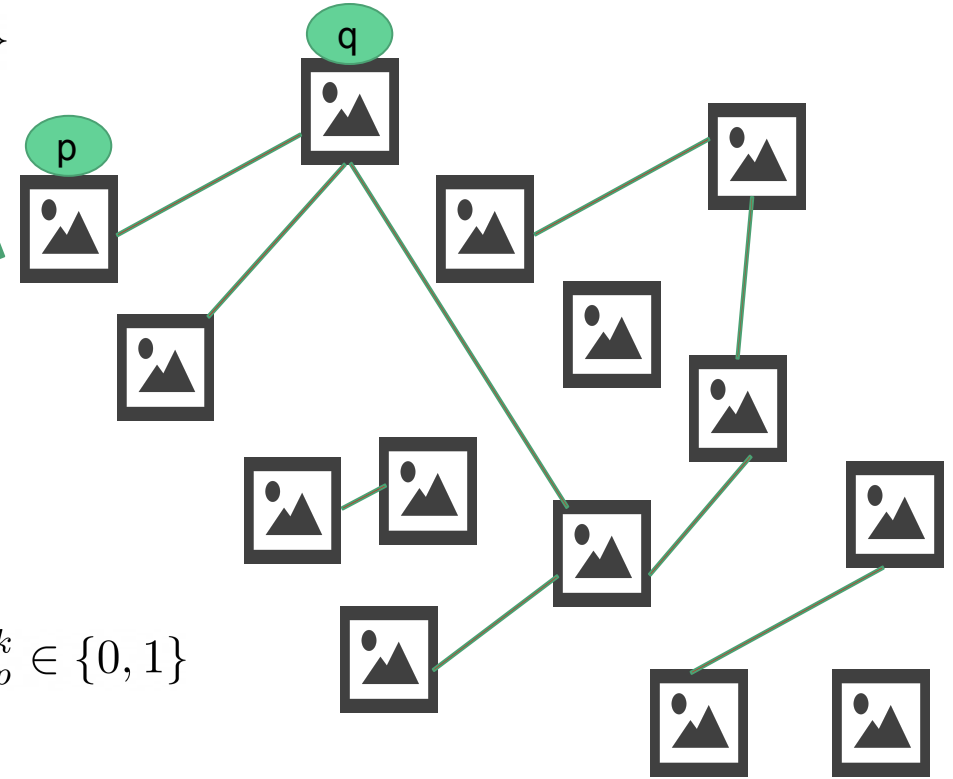


Methodology: Previous work (Vo et al. 2019)

- edge between image p,q : $e_{pq} \in \{0, 1\}$



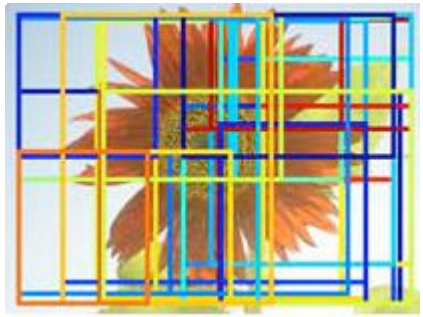
r candidate proposals



- Objectness of k-th proposal in image p: $x_p^k \in \{0, 1\}$

Image source [5]

Methodology: Previous work (Vo et,al 2019)



r candidate proposals

$$\mathbf{x_p} = \begin{bmatrix} x_p^1 \\ x_p^2 \\ \vdots \\ x_p^r \end{bmatrix}$$

$$\mathbf{S_{pq}} = \begin{bmatrix} d(p^1, q^1) & \dots & d(p^1, q^r) \\ \vdots & & \\ d(p^r, q^1) & \dots & d(p^r, q^r) \end{bmatrix}$$

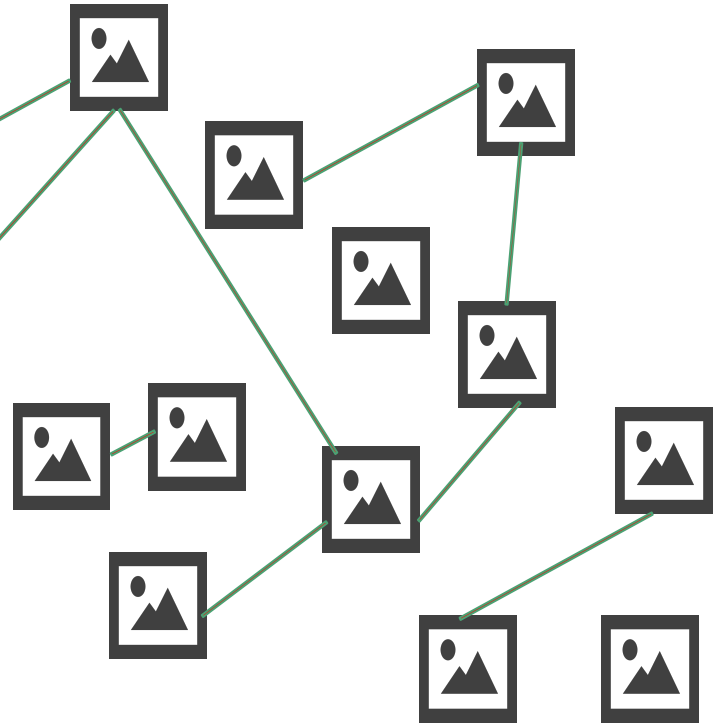


Image source [5]

Methodology: Previous work (Vo et al. 2019)

- edge between image p,q : $e_{pq} \in \{0, 1\}$
- Objectness of k-th proposal in image p: $x_p^k \in \{0, 1\}$
- Edges connect visually similar images

$$\mathbf{x_p} = \begin{bmatrix} x_p^1 \\ x_p^2 \\ \vdots \\ x_p^r \end{bmatrix} \qquad \mathbf{S_{pq}} = \begin{bmatrix} d(p^1, q^1) & \dots & d(p^1, q^r) \\ \vdots & & \\ d(p^r, q^1) & \dots & d(p^r, q^r) \end{bmatrix}$$

$$\max_{x,e} \sum_{p=1}^{n} \sum_{q=1}^{n} e_{pq} \mathbf{x_p^T S_{pq} x_q}$$

$$s.t. \sum_{k=1}^{r} x_p^k \leq \nu \quad and \quad \mathbf{x_p^T} \sum_{q \in N(p)} \sum_{p \neq q} \mathbf{S_{pq}} \mathbf{x_q} \leq \tau$$

$$\max_{x} \sum_{p=1}^{n} \mathbf{x_p^T} \sum_{q \in N(p)} \mathbf{S_{pq} x_q}$$

ν : Maximum #objects in an image

τ : Maximum #neighbors of an image

Methodology: Previous work (Vo et al. 2019)

$$\max_{x,e} \sum_{p=1}^{n} \sum_{q=1}^{n} e_{pq} \mathbf{x_p^T S_{pq} x_q} \qquad s.t \sum_{k=1}^{r} x_p^k \leq \nu \; and \; \sum_{p \neq q} e_{pq} \leq \tau$$

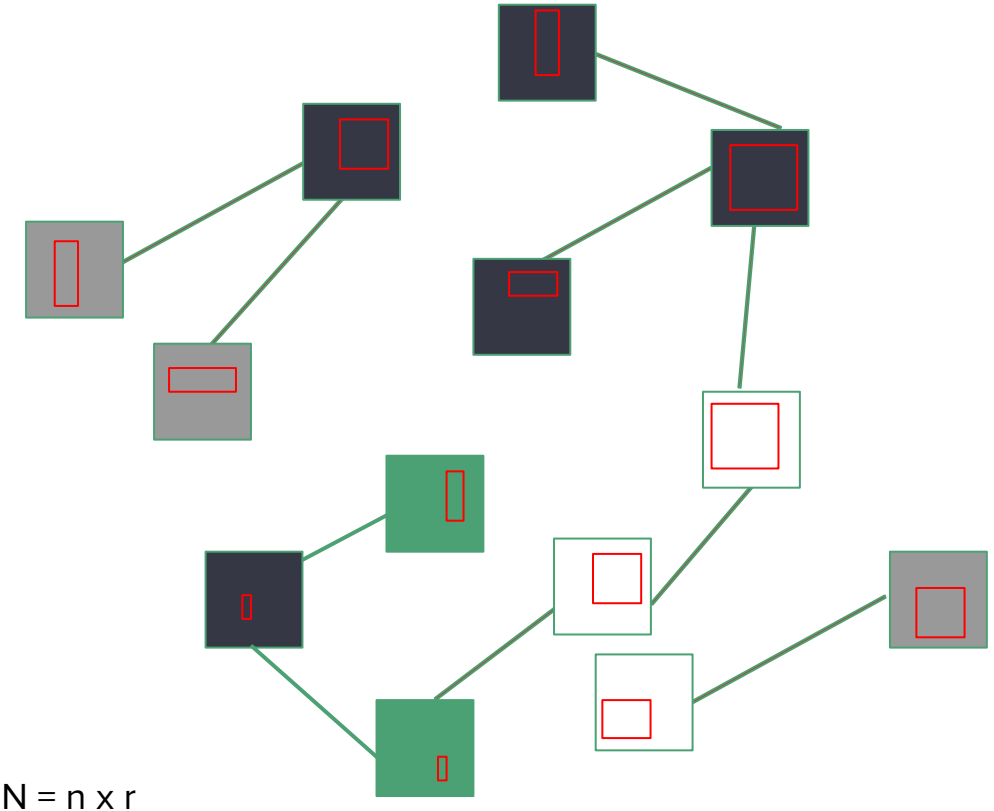
- A relaxed version of this problem is considered
- The solutions are mapped back to the discrete space using an iterative greedy ascent approach

Bottleneck

- A follow up work by Vo et al. 2020 tries to scale this up by reducing r
- But was found to hinder the ability to discover multiple objects

Methodology: This paper (Vo et al. 2021)

- Assume n images, r candidate proposals per image
- Assume Implicit graph structure over Image proposals
- Each node denoted by p^k : p-th image, k-th proposal



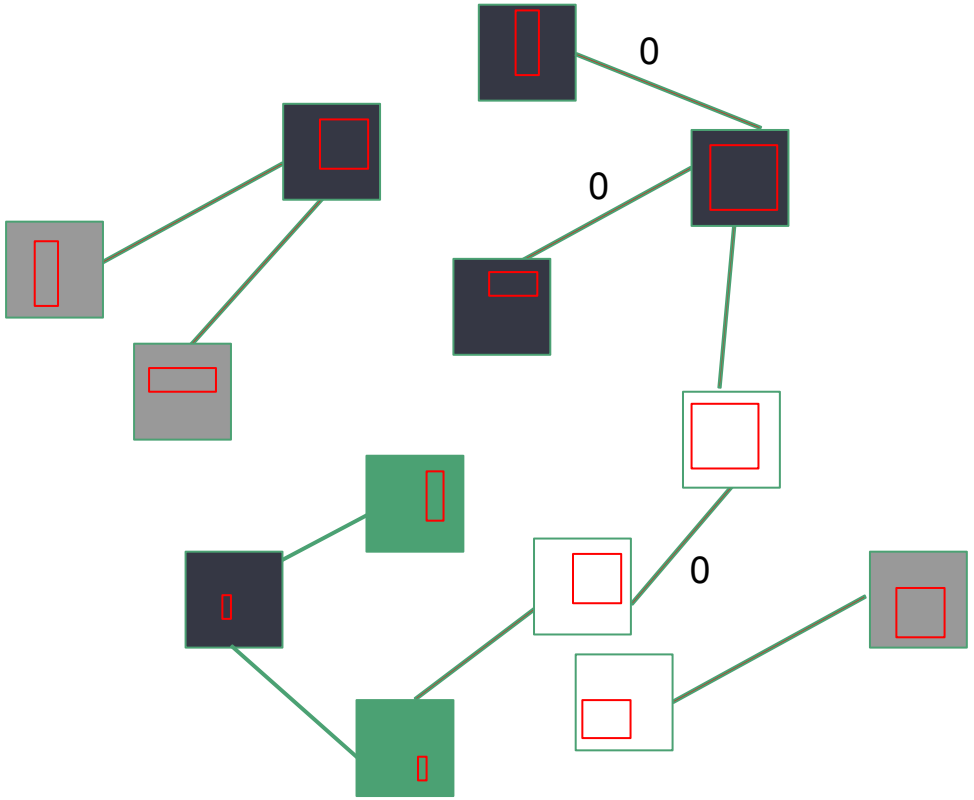
Total nodes N = n x r

Methodology: This paper (Vo et al. 2021)

$$\mathbf{S_{pq}} = \begin{bmatrix} d(p^1, q^1) & \dots & d(p^1, q^r) \\ \vdots & & \\ d(p^r, q^1) & \dots & d(p^r, q^r) \end{bmatrix}$$

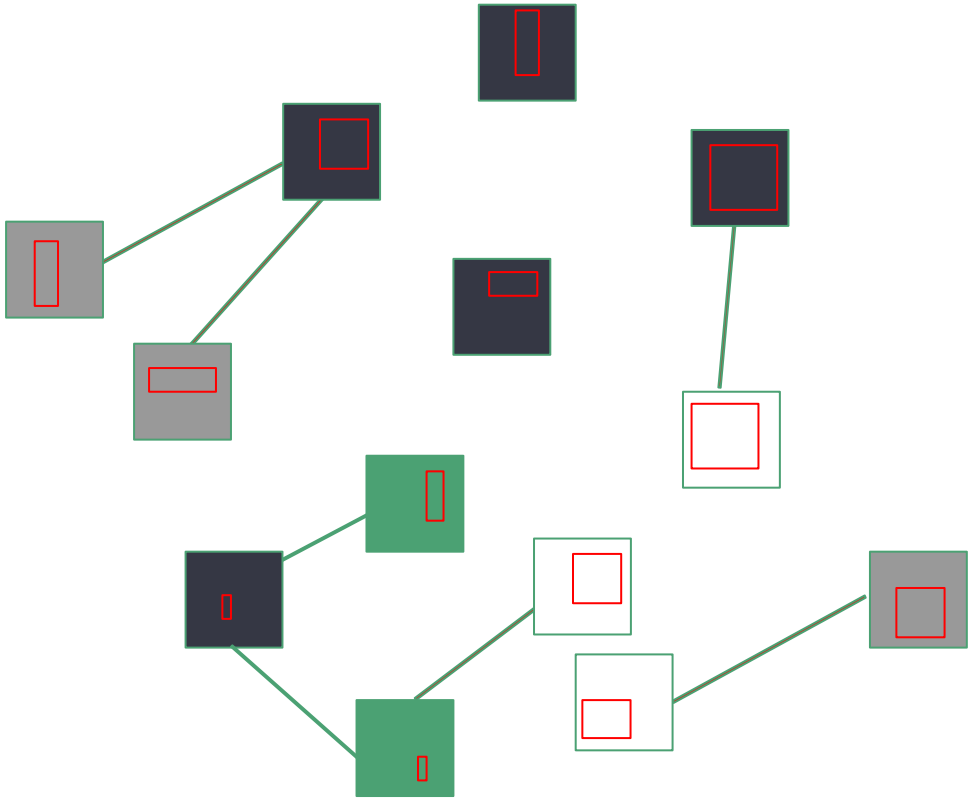
- Two nodes p^k , q^l have edge weight $S_{pq}^{k,l}$

- If p=q, $S_{pq} = 0^{r \times r}$



Methodology: This paper (Vo et al. 2021)

- Goal: Find the proposals with prominent objects
- Idea: A node has a prominent object if its neighbors have prominent objects
- Rank according to an important measure y_i for all nodes i

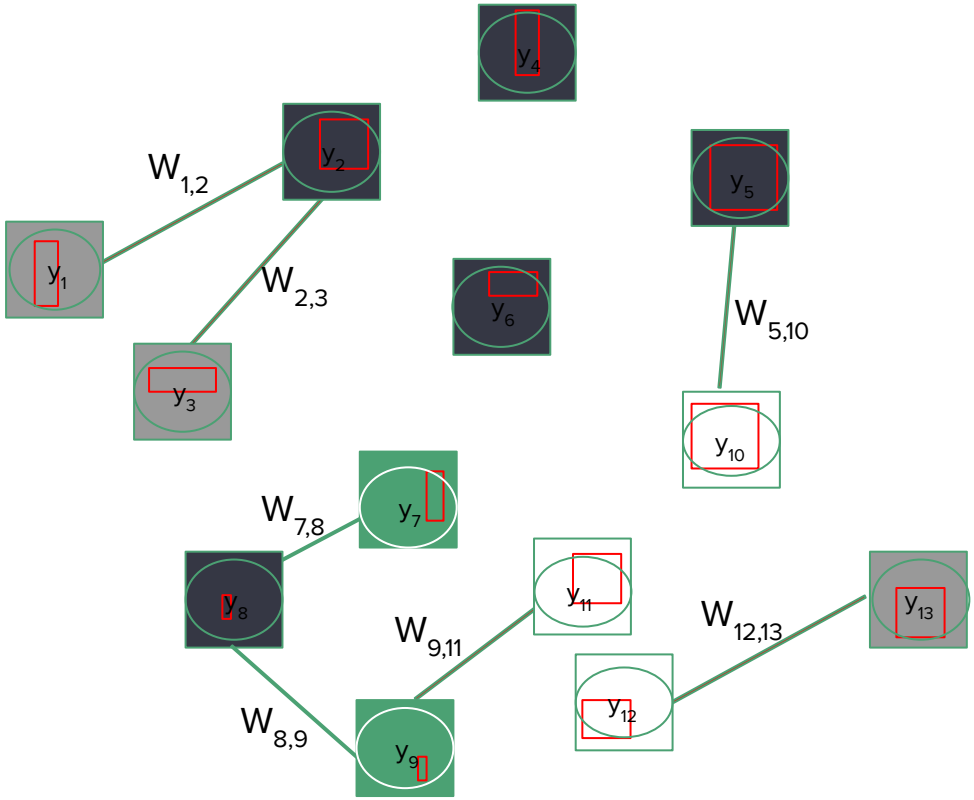


Methodology: This paper (Vo et al. 2021)

- Idea: A node is important if its neighbors are important

- Compute $z_i = \sum_{j \in N(i)} W_{ji}\, y_j$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix}$$



Methodology: This paper (Vo et al. 2021)

- $z_i = \sum_{j \in N(i)} W_{ji}\, y_j \longrightarrow \mathbf{z} = \mathbf{W}\mathbf{y}$ $\qquad \max_{\mathbf{y}} \mathbf{y}^{\mathbf{T}}\mathbf{z} = \max_{\mathbf{y}} \mathbf{y}^{\mathbf{T}}\mathbf{W}\mathbf{y}$

Lemma 1. *Suppose W is irreducible (i.e., represents a strongly connected graph G). The solution y^* of the quadratic optimization problem:*

$$y^* = \underset{\|t\| \leq 1, t \geq 0}{\text{argmax}}\ t^T W t \qquad \text{(Q)}$$

is the unique unit, non-negative eigenvector of W associated with its largest eigenvalue.

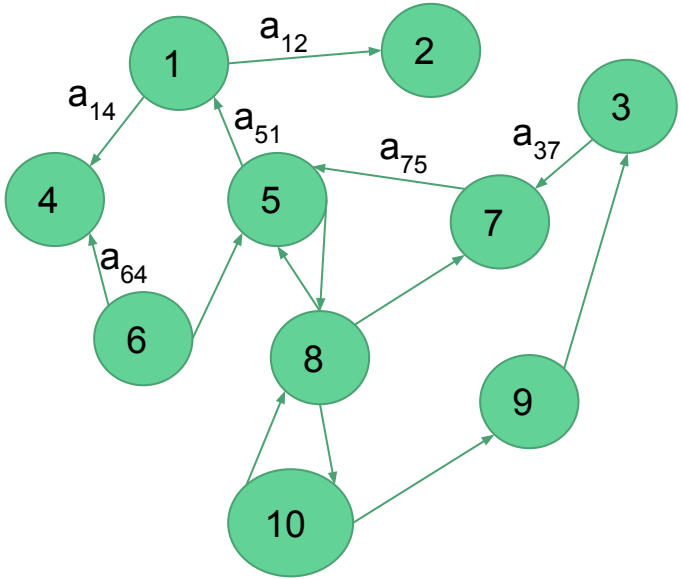
- W is not irreducible (our G is not strongly connected), Add $\frac{\gamma}{N}\begin{bmatrix} 1 & \dots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$ to **W**

Distributed algorithms & tools for eigen value problems can be utilized

Image source [1]

Methodology: Applicability of PageRank (P)

- PageRank
  - Nodes are web-pages
  - Edges are links
  - Goal: Rank the pages based on importance
- Idea
  - a page is important if its backlinks are important
- Transition Matrix $\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1N} \\ \vdots & a_{pq} & \vdots \\ a_{N1} & \dots & a_{NN} \end{bmatrix}$
- Normalize columns in our W to apply PageRank



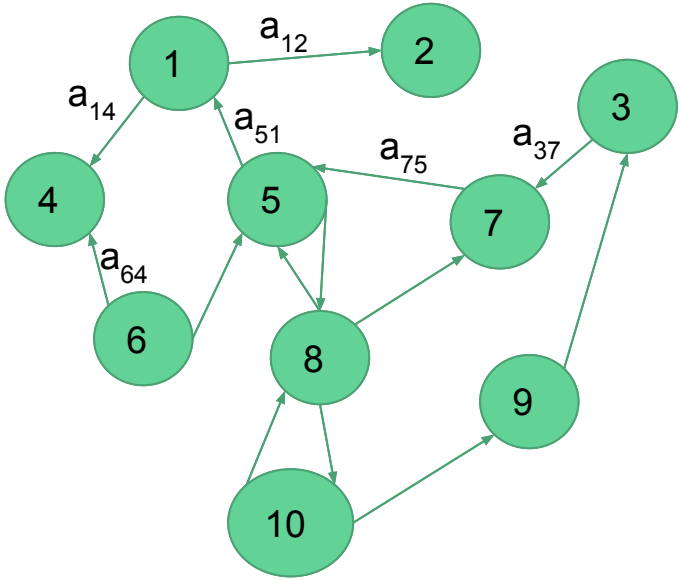
Methodology: Hybrid approach (LOD)

- General form of Transition matrix

$$(1-\beta)A+\beta\begin{bmatrix} u_1 & u_1 & \dots & u_1 \\ u_2 & u_2 & \dots & u_2 \\ \vdots & & & \vdots \\ u_N & u_N & \dots & u_N \end{bmatrix}$$

  u_j importance is manually added to node j

- Compute a maximizer solution **y** using (Q) and then use it as **u** to apply (P)



Outline

- Related works
- Methodology
- **Baselines**
- Experimental Settings
- Results & Conclusion

Baselines (can scale to large datasets)

- EdgeBoxes (EB)
  - Unsupervised method
  - Outputs regions in an image with an importance score
- Kim (2009)
  - candidate ROI set provided for each image
  - Maintain a current ROI set
  - 1st step: find good ROIs (hubs) among current ROI set u pagerank
  - 2nd step: Update current ROI set based on hubs
- Wei (2019)
  - A scalable approach for Image co-localization

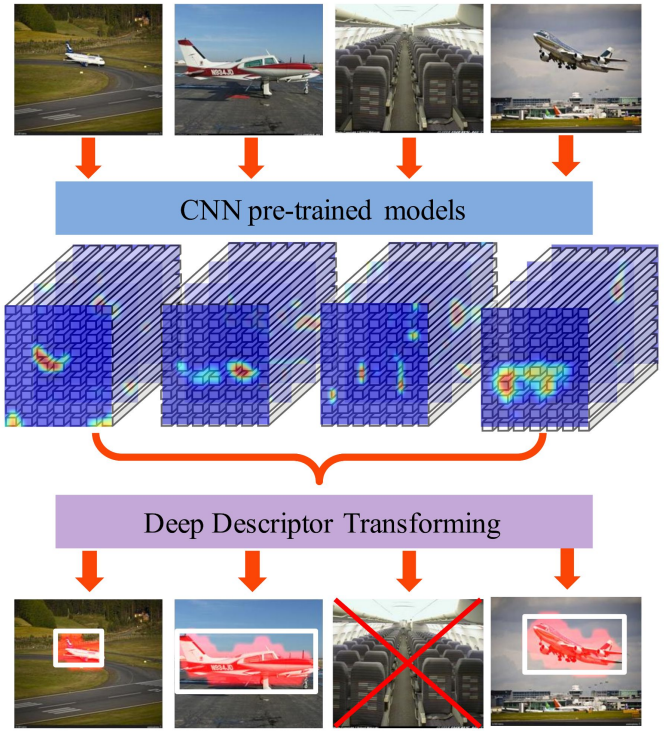


Image source [4]

Outline

- Related works
- Methodology
- Baselines
- **Experimental Settings**
- Results & Conclusion

Experiment Settings

- Datasets

C120K	**OP1.7M**	**C20K**	**OP50K**
Derived from MSCOCO	OpenImages	Subset of C120K	Subset of OP1.7M
~120K Images	~1.7M Images	~20K Images	~50K Images

- Generating Region Proposals
  - Vo (2020) proposed an algorithm which uses activations from pre-trained cnns
  - VGG16 trained on Imagenet with labels (supervised setting)
  - VGG OBoW trained without labels (self-supervised setting)

Experiment Settings

- Similarity Model
  - Probabilistic Hough Matching (PHM) algorithm (Cho et al. 2015)
    - Compares local appearance and global geometric consistency of proposals

- Evaluation Settings
  - Single object discovery
    - Return m=1 region proposal per image
  - Multi object discovery
    - Return m=M region proposals per image
    - M = max #objects in any image

Evaluation Metrics

- Single Object Discovery
  - Correct localization score (CorLoc)
    - Percentage of proposals correctly localized
    - If IOU >= 0.5 between one of the ground truth boxes and predicted proposal

- Multi Object Discovery
  - Average Precision (AP50)
    - Area under PR curve by varying m=1 to M (at 50% IOU threshold)
  - Average Precision (AP@[50:95])
    - Average of AP under different IOU thresholds (10 equal intervals from 50% to 95%)

Outline

- Related works
- Methodology
- Baselines
- Experimental Settings
- **Results & Conclusion**

Results

- Vo [67] is previous work of Vo et al. 2020

Method	Single-object				Multi-object							
	CorLoc				AP50				AP@[50:95]			
	C20K	C120K	Op50K	Op1.7M	C20K	C120K	Op50K	Op1.7M	C20K	C120K	Op50K	Op1.7M
EB [83]	28.8	29.1	32.7	32.8	4.86	4.91	5.46	5.49	1.41	1.43	1.53	1.53
Wei [71]	38.2	38.3	34.8	34.8	2.41	2.44	1.86	1.86	0.73	0.74	0.6	0.6
Kim [32]	35.1	34.8	37.0	-	3.93	3.93	4.13	-	0.96	0.96	0.98	-
Vo [67]	**48.5**	48.5	48.0	47.8	5.18	5.03	4.98	4.88	1.62	1.6	1.58	1.57
Ours (LOD+Self [18])	41.1	42.4	**49.5**	**49.4**	4.56	4.90	6.37	**6.28**	1.29	1.37	1.87	**1.86**
Ours (LOD)	**48.5**	**48.6**	48.1	47.7	**6.63**	**6.64**	**6.46**	**6.28**	**1.98**	**2.0**	**1.88**	1.83

Image source [1]

Results

- Vo [67] is previous work of Vo et al. 2020

Opt.	Proposal	Feature	Single-object CorLoc C20K	Single-object CorLoc Op50K	Multi-object AP50 C20K	Multi-object AP50 Op50K	Multi-object AP@[50:95] C20K	Multi-object AP@[50:95] Op50K
None	EB [83]	None	28.8	32.7	4.86	5.46	1.41	1.53
	[67]+Self		29.7	39.8	2.47	3.72	0.61	1.0
	[67]+Sup		23.6	38.1	4.07	4.81	1.03	1.39
Wei [71]	None	Self	37.9	42.4	2.53	3.13	0.69	0.9
		Sup	38.2	34.8	2.41	1.86	0.73	0.6
Kim [32]	EB [83]	Self	5.5	5.4	0.64	0.79	0.13	0.15
		Sup	15.6	20.2	1.96	2.56	0.36	0.47
	[67]+Self	Self	4.7	4.6	0.13	0.29	0.02	0.05
	[67]+Sup	Sup	35.1	37.0	3.93	4.13	0.96	0.98
Vo [67]	EB [83]	Self	35.6	43.6	3.34	4.43	0.99	1.39
		Sup	40.2	44.0	4.0	4.47	1.21	1.41
	[67]+Self	Self	37.8	48.1	2.65	4.19	0.82	1.45
	[67]+Sup	Sup	**48.5**	48.0	5.18	4.98	1.62	1.58
LOD	EB [83]	Self	35.5	39.7	5.87	6.73	1.57	1.76
		Sup	38.9	41.3	6.52	**7.01**	1.76	1.86
	[67]+Self	Self	41.1	**49.5**	4.56	6.37	1.29	1.87
	[67]+Sup	Sup	**48.5**	48.1	**6.63**	6.46	**1.98**	**1.88**

Opt.	Proposal	Feature	Single-object CorLoc C20K	Single-object CorLoc Op50K	Multi-object AP50 C20K	Multi-object AP50 Op50K	Multi-object AP@[50:95] C20K	Multi-object AP@[50:95] Op50K
Q	EB [83]	Self	32.8	40.3	4.15	6.43	1.07	1.67
		Sup	36.0	41.1	5.72	6.49	1.47	1.7
	[67]+Self	Self	38.7	48.9	4.38	6.39	1.17	1.84
	[67]+Sup	Sup	43.8	47.5	6.21	6.66	1.74	**1.88**
P	EB [83]	Self	35.5	39.7	4.91	6.73	1.34	1.75
		Sup	38.9	41.3	6.51	6.99	1.76	1.86
	[67]+Self	Self	41.2	**49.5**	4.38	6.13	1.24	1.81
	[67]+Sup	Sup	47.5	47.8	6.25	6.19	1.87	1.81
LOD	EB [83]	Self	35.5	39.7	5.87	6.73	1.57	1.76
		Sup	38.9	41.3	6.52	**7.01**	1.76	1.86
	[67]+Self	Self	41.1	**49.5**	4.56	6.37	1.29	1.87
	[67]+Sup	Sup	**48.5**	48.1	**6.63**	6.46	**1.98**	**1.88**

Image source [1]

Results

- Runtime Comparison

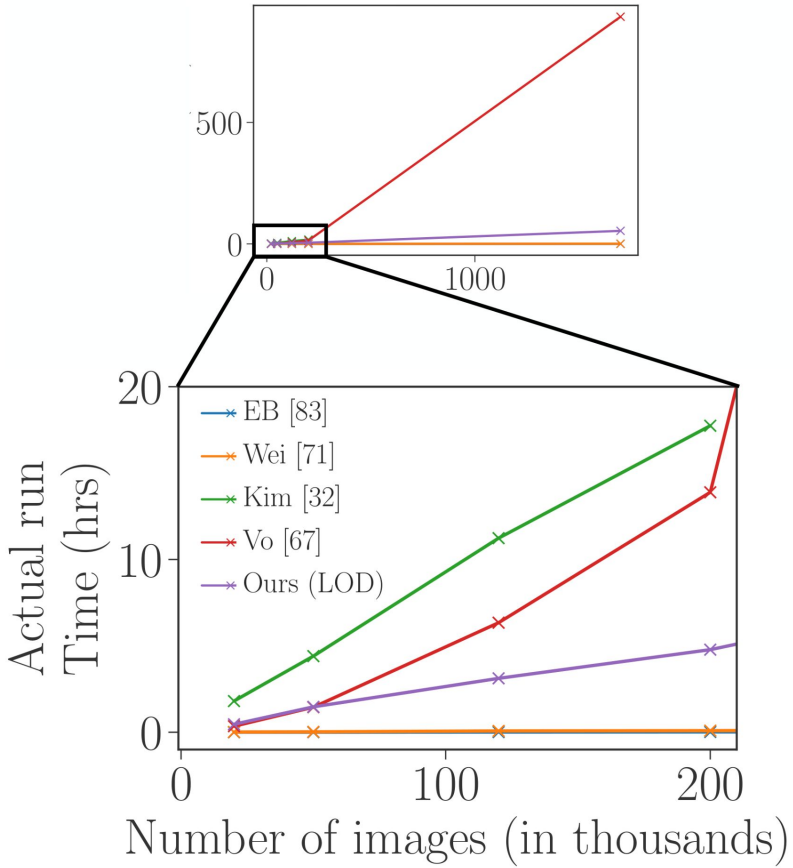


Image source [1]

Results

- Results by LOD on OP1.7M, Ground truth objects in yellow, predictions in red

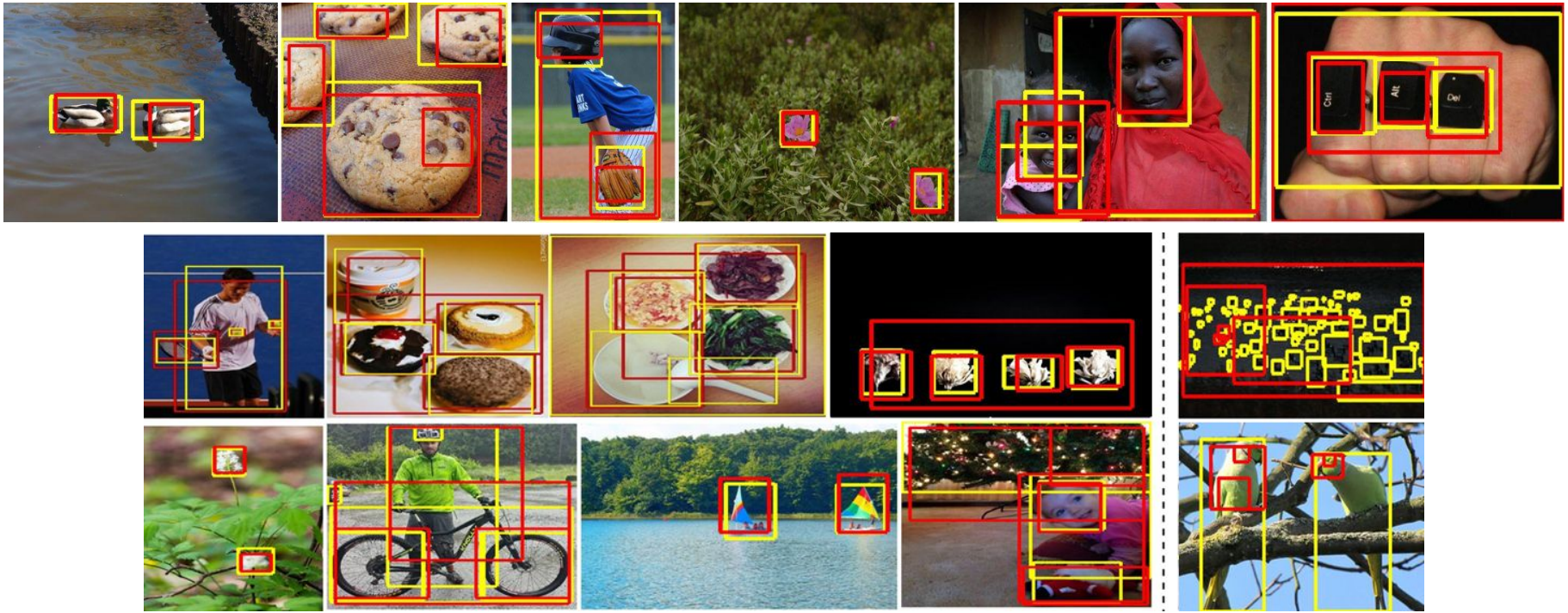


Image source [1]

Summary

- Novel formulation of UOD as ranking problem
- LOD can be scaled to very large datasets (eg. OP1.7M)
- LOD + self-supervised features (state of the art on OP1.7M, Multi object setting)
- Only works well with VGG16 features
- Small objects are not discovered well
- LOD is found to be sensitive to hyperparameters

References

[1] Vo, Van Huy, et al. "Large-scale unsupervised object discovery." *Advances in Neural Information Processing Systems* 34 (2021): 16764-16778.

[2] Vo, Huy V., Patrick Pérez, and Jean Ponce. "Toward unsupervised, multi-object discovery in large-scale image collections." *ECCV 2020-16th European Conference on Computer Vision*. 2020.

[3] Vo, Huy V., et al. "Unsupervised image matching and object discovery as optimization." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[4] Wei, Xiu-Shen, et al. "Unsupervised object discovery and co-localization by deep descriptor transformation." *Pattern Recognition* 88 (2019): 113-126.

[5] Kim, Gunhee, and Antonio Torralba. "Unsupervised detection of regions of interest using iterative link analysis." *Advances in neural information processing systems* 22 (2009).

[6] Cho, Minsu, et al. "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[7] Tuytelaars, Tinne, et al. "Unsupervised object discovery: A comparison." *International journal of computer vision* 88.2 (2010): 284-302.

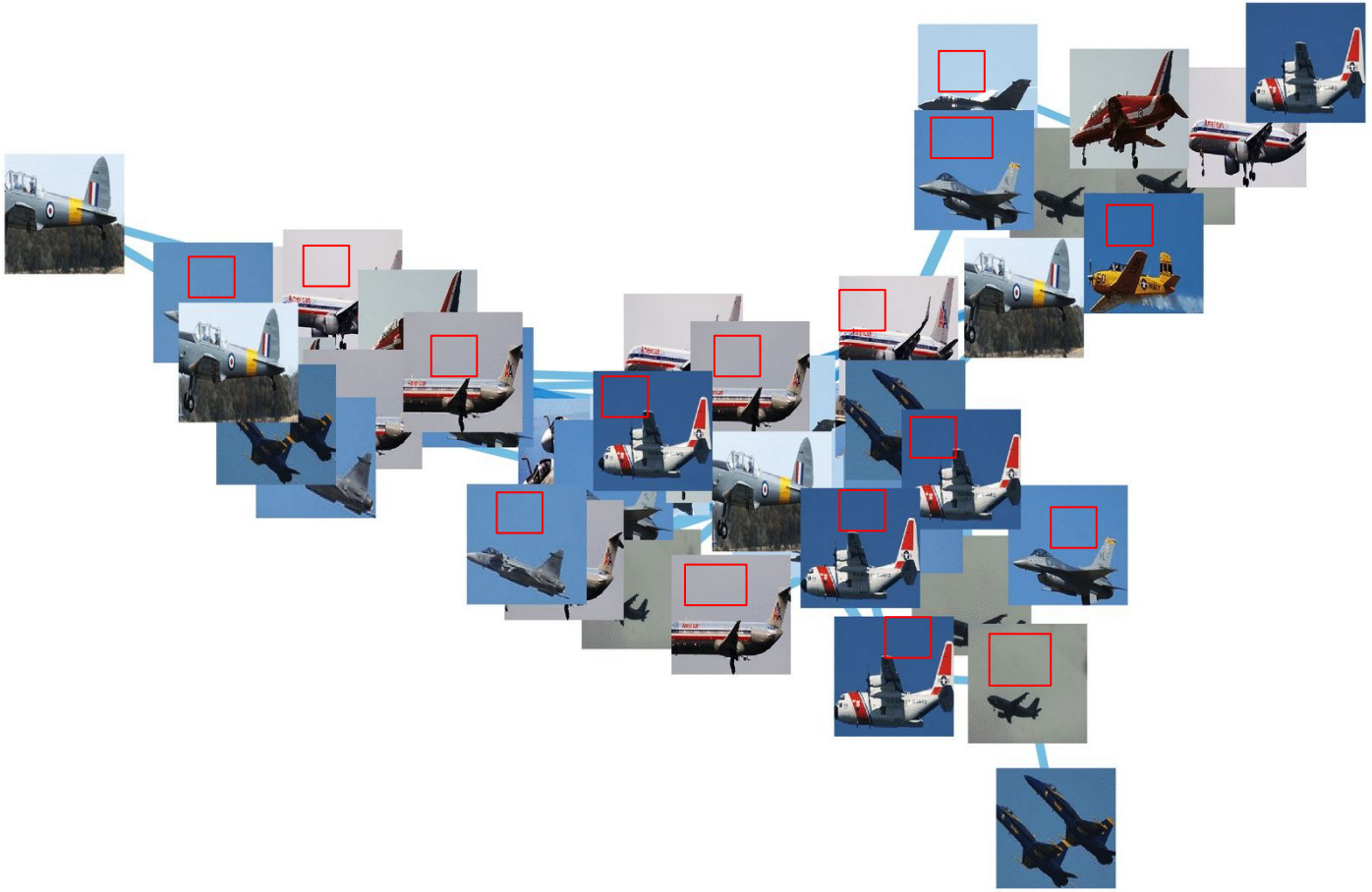
Ur

Unsupervised Object Discovery

- Detect prominent objects in Large collection of images
- Ground truth bounding boxes are considered true objects
- Challenge: prominent objects could be some meaningless textures

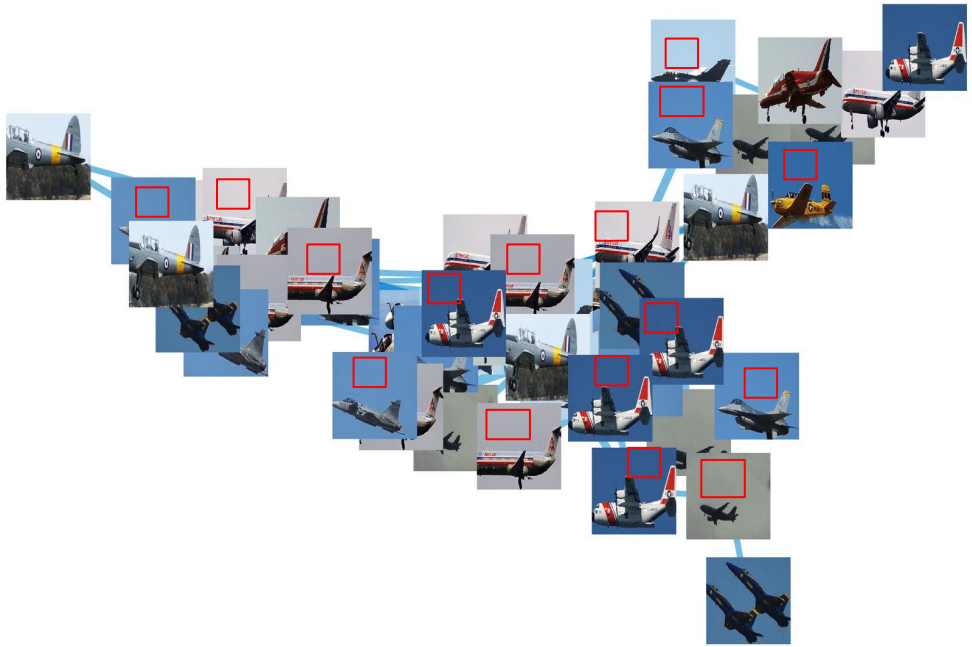
Table 1: Large-scale object discovery performance and comparison to the state of the art on COCO [42] (C120K), OpenImages [35] (Op1.7M) and their respective subsets C20K and Op50K, in three standard metrics. Using VGG16 features [60], the proposed method LOD achieves top performance in both single and multi-object discovery, and scales better to 1.7M images in Op1.7M than the previous state of the art [67]. When running with self-supervised features (LOD + Self [18]), it yields the best results on Op1.7M, showing the first effective fully unsupervised pipeline for UOD. See Sec. 4 for more details.

Method	Single-object				Multi-object							
	CorLoc				AP50				AP@[50:95]			
	C20K	C120K	Op50K	Op1.7M	C20K	C120K	Op50K	Op1.7M	C20K	C120K	Op50K	Op1.7M
EB [83]	28.8	29.1	32.7	32.8	4.86	4.91	5.46	5.49	1.41	1.43	1.53	1.53
Wei [71]	38.2	38.3	34.8	34.8	2.41	2.44	1.86	1.86	0.73	0.74	0.6	0.6
Kim [32]	35.1	34.8	37.0	-	3.93	3.93	4.13	-	0.96	0.96	0.98	-
Vo [67]	**48.5**	48.5	48.0	47.8	5.18	5.03	4.98	4.88	1.62	1.6	1.58	1.57
Ours (LOD+Self [18])	41.1	42.4	**49.5**	**49.4**	4.56	4.90	6.37	**6.28**	1.29	1.37	1.87	**1.86**
Ours (LOD)	**48.5**	**48.6**	48.1	47.7	**6.63**	**6.64**	**6.46**	**6.28**	**1.98**	**2.0**	**1.88**	1.83



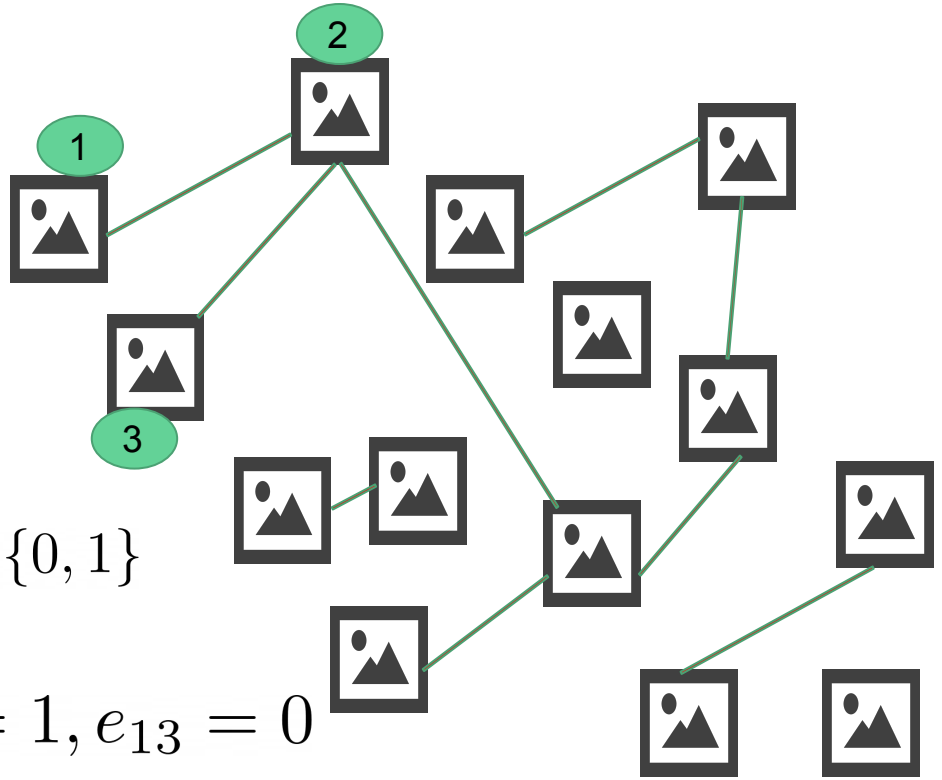
Unsupervised Object Discovery

- Challenge: prominent objects could be some meaningless textures
- Measures for foregroundness, standout score etc
- Pre-trained cnn activations contain cues to locate objects



Methodology: Previous work (Vo et al. 2019)

- Initial work on modeling UOD as an optimization problem
- Assume Implicit graph structure over Images
- edge between image p,q : $e_{pq} \in \{0, 1\}$



$$e_{12} = 1, e_{23} = 1, e_{13} = 0$$

Experiment Settings

- Datasets
  - C120K (derived from MSCOCO), C20K
  - OpenImages (OP1.7M), OP50K
- Generating Region Proposals
  - Vo (2020) proposed an algorithm which uses activations from pre-trained cnns
  - VGG16 trained on Imagenet with labels (supervised setting)
  - VGG OBoW trained without labels (self-supervised setting)
- Similarity Model
  - Probabilistic Hough Matching (PHM) algorithm
- Evaluation Settings
  - Single object discovery: Return m=1 region per image
  - Multi object discovery: Returns m=M regions per image (M = max #objects in any image)