The Pursuit of Knowledge:
Discovering and Localizing Novel Categories using Dual Memory

Sai Saketh Rambhatla[1] Rama Chellappa[2] Abhinav Shrivastava[1]
[1]University of Maryland, College Park [2]Johns Hopkins University

Contents

Contents

1. Implementation Details and Pseudo Codes

In this section, we provide the implementation details of training the classifiers and pseudo codes for the Merge (Alg. 3), Refine (Alg. 1) operations and the overall pipeline (Alg. 2) described in the main paper.

1.1. Pseudo Codes

Algorithm 1 Refine Clusters (refine_clusters)

```
Input: $\mathcal{W}$, $\boldsymbol{b}$, Memory $\mathcal{M}$, Clusters $\boldsymbol{\mathcal{C}}$, number of clusters $m$
for $k \in \{1, 2, \cdots, m\}$ do
    for $f_k$ in $\boldsymbol{\mathcal{C}}[k]$ do
        if $\mathcal{W}[k,:]^T f_k + b[k] < 0$ then
            Remove sample from $\boldsymbol{\mathcal{C}}[k]$
            $c_k = \text{len}(\boldsymbol{\mathcal{C}}[k])$
            $\mathcal{M}_w[k,:] = \frac{(c_k * \mathcal{M}[k,:] - f_k)}{c_k - 1}$; $c_k = c_k - 1$
        end if
    end for
end for
return $\mathcal{W}$, $\boldsymbol{b}$, $\mathcal{M}$
```

Algorithm 2 Overall Pipeline

```
Input: Discovery Set $\mathcal{D}$ containing known object categories $\mathcal{O}_k$; covariance matrix $\Sigma$, mean $\mu_0$; Storage $\mathcal{S}$ containing Semantic, Working memory ($\mathcal{M}_s$, $\mathcal{M}_w$).
Initialization: $\mathcal{M}_s \leftarrow$ semantic prior
for True do
    for $\mathcal{I} \in \mathcal{D}_1$ do
        $f, r = \text{encode}(\mathcal{I})$ // Extract features and ROIs
        flag $= \text{retrieve}(\mathcal{S}, f)$
        update_or_create(flag, $\mathcal{M}_s$, $\mathcal{M}_w$, f)
    end for
    filter($\mathcal{M}_w$) // Remove small clusters
    merge_clusters($\mathcal{M}_w$, $\mathcal{W}_w$, $\mathbf{b}_w$)
    refine_clusters($\mathcal{M}_w$, $\mathcal{W}_w$, $\mathbf{b}_w$)
    refine_clusters($\mathcal{M}_s$, $\mathcal{W}_s$, $\mathbf{b}_s$)
    for $\mathcal{I} \in \mathcal{D}_2$ do
        $f, r = \text{encode}(\mathcal{I})$ // Extract features and ROIs
        flag $= \text{retrieve}(\mathcal{S}, f)$
        if flag == 'in $\mathcal{M}_s$' then
            update($\mathcal{M}_s$, f)
        end if
    end for
    merge_clusters($\mathcal{M}_s$, $\mathcal{W}_s$, $\mathbf{b}_s$)
    refine_clusters($\mathcal{M}_s$, $\mathcal{W}_s$, $\mathbf{b}_s$)
    $\mathcal{D}_1, \mathcal{D}_2 = \mathcal{D}_2, \mathcal{D}_1$
end for
```

Algorithm 3 Merge Clusters (merge_clusters)

```
Input: $\mathcal{W}$, $\boldsymbol{b}$, Memory $\mathcal{M}$, Cluster $\boldsymbol{\mathcal{C}}$,
while rounds $< r_{\text{max}}$ do
    $w^j(x) = \mathcal{W}[j,:]^T x + b[j]$
    $A = [a_{ij}] = \frac{1}{2}(\max(0, w^i(\mathcal{M}[j,:])) + \max(0, w^j(\mathcal{M}[i,:])))$
    $D = \text{diag}(A \times \mathbf{1})$
    $L = D - A$
    $[V, E] = eig(L)$
    $n = \text{where}(E < 0.01)$
    $N = \text{kmeans}(V[:, n], len(n))$
    for cluster $c \in$ N do
        if len$(c) > 1$ then
            $c_a \leftarrow$ largest cluster in $c$ with classifier $(\mathcal{W}_a, \mathbf{b}_a)$
            for $c_n \in c \setminus \{c_a\}$ do
                for $x \in c_n$ do
                    affinity $a = (\mathcal{W}_a^T x + \mathbf{b}_a) - (\mathcal{W}_n^T x + \mathbf{b}_n)$
                    if affinity $> 0$ then
                        $c_n \leftarrow c_n \setminus \{x\}$;$c_a \leftarrow c_a \cup \{x\}$
                    end if
                end for
            end for
        end if
    end for
end while
return $\mathcal{W}$, $\boldsymbol{b}$
```

1.2. Estimation of Σ and μ_0

We extract $N = 150$ boxes and their corresponding classification head features after ROI [1] pooling from each image. We use an online batch update equation to estimate Σ and μ_0. Let $\Sigma_{\text{old}}, \mu_0^{\text{old}}$ be the estimated covariance matrix and mean respectively using m samples and $\Sigma_{\text{current}}, \mu_0^{\text{current}}$ be the covariance matrix and mean for the N samples. The covariance matrix and mean for $m + N$ samples can be estimated as follows

$$\hat{\Sigma}_{\text{old}} = \frac{m-1}{m+N-1}\hat{\Sigma}_{\text{old}} + \frac{N-2}{m+N-1}\hat{\Sigma}_{\text{current}} + \frac{m * N}{(m+N)*(m+N-1)} \times (\mu_0^{\text{old}} - \mu_0^{\text{current}})(\mu_0^{\text{old}} - \mu_0^{\text{current}})^T \tag{1}$$

$$\mu_0^{\text{old}} = \frac{m * \mu_0^{\text{old}} + N * \mu_0^{\text{current}}}{m+N} \tag{2}$$

$$m \leftarrow m + N \tag{3}$$

This step needs to be performed only once before the discovery process.

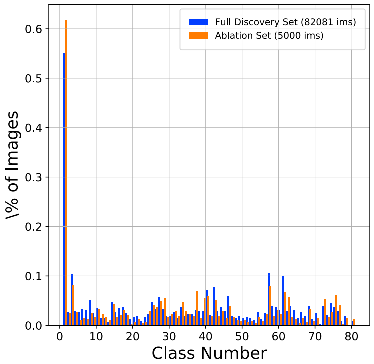

Figure 1: Fraction of images per class in the discovery and ablation sets.

2. Classes removed from ImageNet

In Table 1, we show classes removed from ImageNet for every novel class in COCO dataset.

3. Training details

Training backbone. We train a ResNet-101 [2] model on the 932 ImageNet$^-$ classes, using stochastic gradient descent (SGD) with an initial learning rate, momentum, weight decay of 0.1, 0.9, 1×10^{-4}, respectively and a mini-batch size of 256 on 8 GPUs. We train for 90 epochs and decrease the learning rate by a factor of 10 every 30 epochs. This model achieves top-1/5 accuracy of 78.12/93.97% on the ImageNet$^-$ val set. Comparison of this to the same model trained on ImageNet, which gets 76.4/92.9% on ImageNet val set, ensuring that these models are reasonably similar. Additional details are provided in the supplementary section.

Training *known* detector. Using ImageNet$^-$ model weights as initialization, we train a ResNet-101 Faster R-CNN on VOC 2007 train set for 20 known classes (*known* detector), using SGD for 10 epochs with a minibatch size of 24 on 4 GPUs. We use an initial learning rate of 0.01 and decrease it by a factor of 10 after 8 epochs. This ImageNet$^-$-pretrained model achieves 72.03 mAP on the VOC 2007 test set (compare to ImageNet-pretrained model, which gets 74.44 mAP).

4. Details about the Ablation Set

We perform all the experiments in Section 5.4 of the main paper on a random subset of 5000 images. For every object in the dataset, we plot the fraction of images in the discovery (82081 images) and ablation (5000 images) set and show it in Fig. 1. We can see that the distribution of the training and ablation sets are very similar.

Table 1: Classes removed from ImageNet for each novel class in COCO.

Novel Class	**ImageNet Class**	**Novel Class**	**ImageNet Class**	**Novel Class**	**ImageNet Class**	**Novel Class**	**ImageNet Class**
toilet	toilet seat	teddy bear	teddy	kite	kite	stop sign	street sign
tennis racket	racket	snowboard	–	carrot	–	zebra	zebra
oven	microwave	keyboard	computer keyboard typewriter keyboard	scissors	–	fire hydrant	–
mouse	mouse	clock	analog clock digital clock wall clock	frisbee	–	apple	–
hair drier	hand blower	cup	measuring cup & cup	traffic light	traffic light	toaster	toaster
bowl	mixing bowl;soup bowl	microwave	microwave	bench	park bench	book	bookcase;comic book
orange	orange	elephant	Indian elephant African elephant	tie	bolo tie bow tie Windsor tie	banana	banana
knife	letter opener	pizza	pizza	fork	forklift	sandwich	–
umbrella	umbrella	bear	koala bear & brown bear American black bear ice bear & sloth bear giant panda	vase	vase	toothbrush	–
spoon	wooden spoon	giraffe	–	sink	–	wine glass	beer glass
handbag	–	cell phone	cellular telephone dial telephone pay-phone	broccoli	broccoli	refrigerator	refrigerator
laptop	laptop	remote	remote control	surfboard	–	hot dog	hotdog
baseball bat	–	sports ball	baseball & basketball croquet ball & golf ball ping-pong ball & rugby ball soccer ball & tennis ball volleyball	skateboard	–	bed	studio couch
donut	–	truck	fire engine & garbage truck pickup & tow truck trailer truck	skis	–	parking meter	parking meter
suitcase	–	cake	–	baseball glove	–	backpack	backpack

Table 2: Number of discovery iterations.

Rounds	**CorLoc**	**CorRet**	**AuC**		**#disc. objs**
			@0.5	@0.2	
1	40.09	66.22	3.84	10.05	41
2	42.21	66.07	4.27	10.97	45
3	42.52	66.09	4.47	11.27	49
4	42.66	66.39	4.53	11.27	47

Table 3: Discovery threshold

Threshold	**CorLoc**	**CorRet**	**AuC**		# objs
			@0.5	@0.2	
0.50	42.73	48.60	0.80	3.26	3
0.55	42.73	51.76	1.17	3.21	5
0.60	42.73	53.12	1.86	5.45	14
0.65	42.73	56.24	2.91	7.86	26
0.70	42.57	63.12	3.70	9.60	44
0.75	More than 2000 clusters				

5. Extensive Literature Survey

Object Discovery: Hsu et al. [3–5] utilize prior knowledge to facilitate the discovery of image categories. They assume availability of a source dataset with categorical labels and learn a network to predict semantic similarity in a class-agnostic way. They transfer the learnt knowledge, to learn a clustering network on the target dataset with supervision provided by the similarity network. One major concern with such approaches is that the transferred simi-

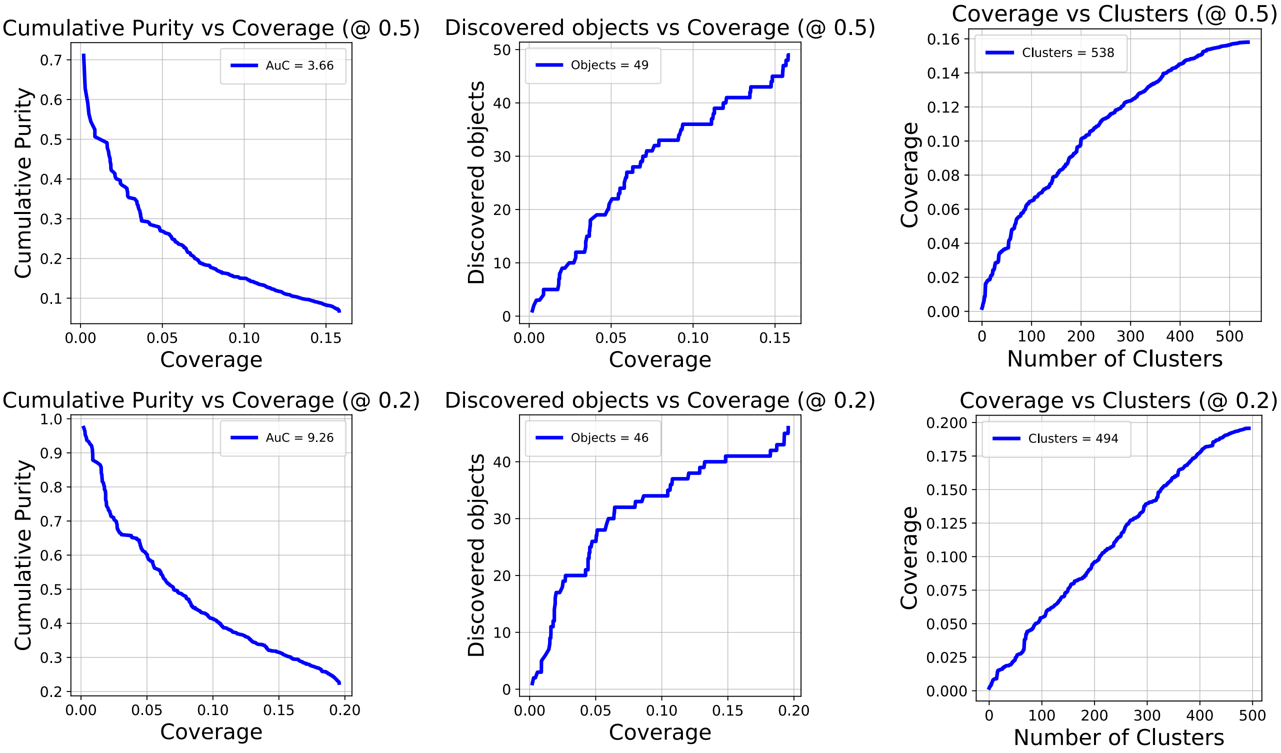

Figure 2: Figure showing the plots of Cumulative Purity vs Coverage, Discovered objects vs Coverage and Coverage vs Number of clusters. These plots are helpful for comparison across future discovery methods. In each case the higher the curve the better. The plots in top row are evaluated with an IoU threshold of 0.5 and the bottom row with a threshold of 0.2.

larity function might produce noisy predictions in the target dataset. Hsu et al. [3–5] posit that dense pair-wise similarity constraints can greatly alleviate this issue. For this method, the number of unknown classes/clusters need to be known a-priori.

Han et al. [6] follow a similar pipeline as above. At its core, their method is based on a deep clustering algorithm that clusters data while learning a good data representation. They learn a feature extractor on the labelled data, and transfer this knowledge by fine-tuning on the unlabeled data and clustering the samples. They also provide a method to determine the number of clusters.

Arandjelović et al. [7] and Singh et al. [8] proposed GAN-based methods for Object Discovery. Arandjelović et al. [7] trained a GAN to segment an object and paste it on another image with the rationale that for the generator to learn and discover objects, it should be able to fool a discriminator by copying the objects into appropriate backgrounds. Singh et al. [8] proposed FineGAN, a novel unsupervised GAN framework, which disentangles the background, object shape, and object appearance to hierarchically generate images of fine-grained object categories. They demonstrate the prowess of their unsupervised disentanglement by using the features from the FineGAN to cluster real world images and discover concepts.

Lee et al. [9] improved upon their previous work by proposing an iterative procedure for category discovery. They proposed a self-paced approach that instead focuses on the easiest instances first, and progressively expands its repertoire to include more complex objects. At each cycle of the discovery process, they re-estimate the easiness of each subwindow in the pool of unlabeled images, and then retrieve a single prominent cluster from among the easiest instances. While our proposed method is also iterative, we do not define any notion of "easiness". We instead let the frequency of occurrence and visual saliency of an object decide when it is learnt in the pipeline.

Doersch et al. [10] leverage context as supervision to discovery visually consistent clusters. The proposed approach gradually discovers visual object clusters together with a segmentation mask.

Kang et al. [11] leveraged multiple segmentations to process noisy clusters and extracted object models as groups of mutually consistent segments. They achieved this by enforcing constraints on geometry (scale, orientation) and appearance (color, texture and shape).

Wang et al. [12] formulated the unsupervised object category discovery as a sub-graph mining problem from a weighted graph of object proposals, where nodes correspond to object proposals, and edges represent the similarities between neighbouring proposals. The objects are discovered by finding sub-graphs of strongly connected nodes, with each sub-graph capturing one object pattern. Authors proposed a maximal-flow based algorithm to solve the graph mining in an efficient manner. Our method is not formulated as an operation on a graph. One of the steps to improve our object clusters leverages Spectral Clustering [13–15] to merge clusters representing similar objects.

Xie et al. [16] proposed a method to provide dense segmentation masks for object discovery in videos. They considered anything which moves as a foreground object in their work and utilized RGB and Optical Flow features as inputs. They linked pixels using a pixel linking criteria and applied a Pixel Trajector RNN (PT-RNN) to learn per-pixel representations for foreground clustering. The system is trained end-to-end using a multi-loss setup. Unlike Xie et al. [16] we attempt to discover objects from images rather than videos which is a slightly difficult task given the unavailability of motion as supervision.

Sivic et al. [17] successfully applied probabilistic Latent Semantic Analysis (pLSA), developed in statistical text literature, for topic discovery in a corpus to category discovery. They leveraged a visual analogue of a word formed by vector quantizing SIFT [18] like region features. Using the proposed method they found object categories and approximate spatial position for a small set of objects.

Incremental Learning: Incremental learning methods learn object models of novel categories in a sequential manner without compromising performance on the existing object categories. However they assume the availability of labelled data for learning the new category. [19] proposed a novel approach, called 'Learning without Memorizing (LwM)', to preserve the information about existing categories, without storing any of their data, while making the classifier progressively learn the new classes. Authors demonstrated that penalizing the changes in classifiers' attention maps helps retain information of the existing classes, as new classes are added, thereby preserving its performance on existing categories.

[20] proposed an Incremental learning method which leverages dual memory to alleviate catastrophic forgetting in image classification. The first memory stores exemplars of existing classes while the second memory is used to store the class statistics of existing classes learnt during the initial training. While the current work is not directly related to incremental learning, IL methods can be employed to learn powerful models once an oracle (human/machine) provides labels for the clusters obtained using our method.

Open World Recognition: [21] equip open world learning models with incremental capabilities to evaluate classification models in a more realistic settings. They evaluated their model by incrementally increasing their vocabulary of known categories and simultaneously evaluate on known and novel categories. While current work is an open world problem, we do not assume labels for data available incrementally.

Never Ending Learning: [22] proposed a webly supervised fully automated method to learn all possible variations of a concept. Their approach leverages vast resources of online books to discover the vocabulary of variance, and intertwines the data collection and modeling steps to alleviate the need for explicit human supervision in training object models. NEIL [23] is a constrained semi-supervised learning (SSL) system that exploits the big scale of visual data to automatically extract common sense relationships and then uses these relationships to label visual instances of existing categories. Authors aimed to build the world's largest visual structured knowledge base with minimum human effort, that reflects the factual content of the images on the Internet. Unlike these methods, we aim to discover and localize visual concepts in a fully unsupervised manner.

6. Additional Experiments

Discovery Rounds: While our approach can be run in a never-ending fashion, to understand how many rounds are needed for this dataset we run our method for multiple rounds and show the results in Table 2. As the number of rounds increase, our approach discovers more objects with an increasing AuC. However, after three rounds, the performance saturates with no increase in number of discovered objects. Therefore, in our main experiments, we run our method for three rounds and report the results.

Discovery Threshold: During the retrieval operation a feature is assigned to a cluster in $\mathcal{M}_w$ based on a threshold on the cosine similarity of the feature with the centroid of the cluster. This threshold indirectly controls the number of clusters and thereby, the number of objects discovered by our pipeline. In Table 3 we perform discovery with various thresholds. We run our full pipeline with a threshold of 0.7.

7. Evaluation Plots

To facilitate comparison with future discovery methods, in Fig. 2 we provide the plots of Cumulative Purity (column-1) and Number of discovered objects (column-2) vs Coverage and Coverage vs # of clusters (column-3) for IoU thresholds of 0.5 and 0.2 (top and bottom) respectively.

8. Qualitative Results

In this section we show examples of clusters obtained from our method. Kindly refer to Figures 3, 4, 5, 6, 7, 8, 9.

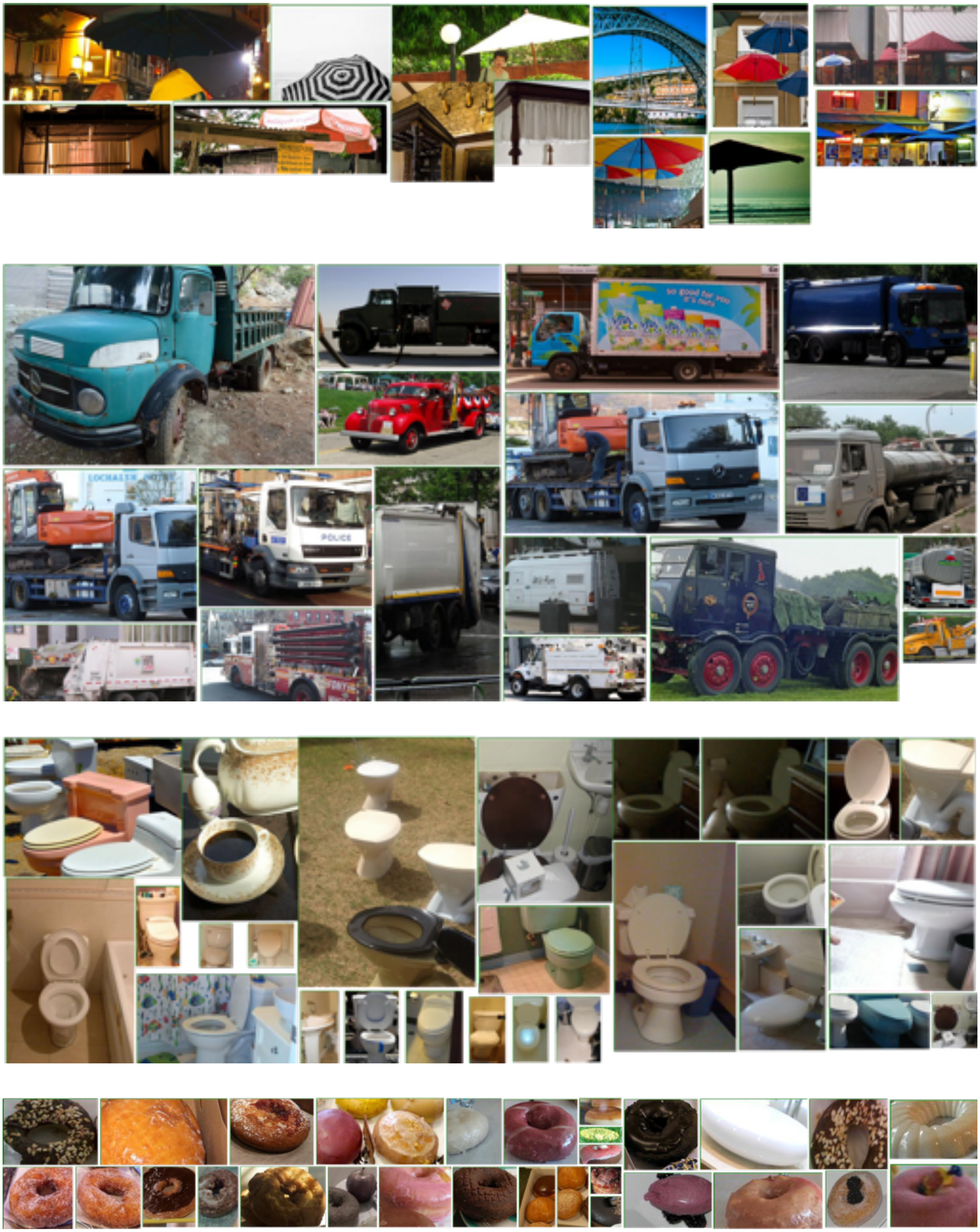

Figure 3: **Qualitative results**: Clusters discovered by our approach.

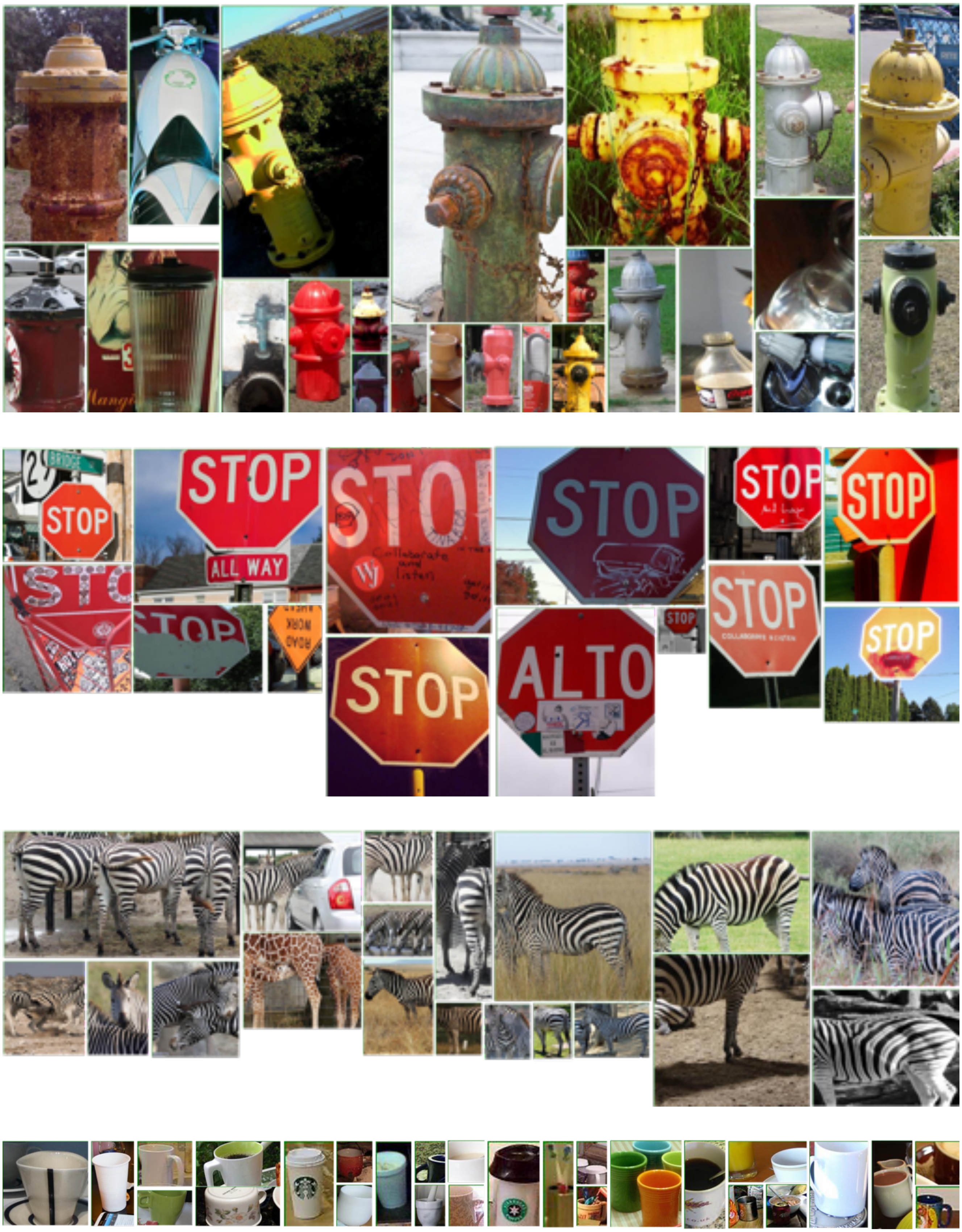

Figure 4: **Qualitative results (cont.)**: Clusters discovered by our approach.

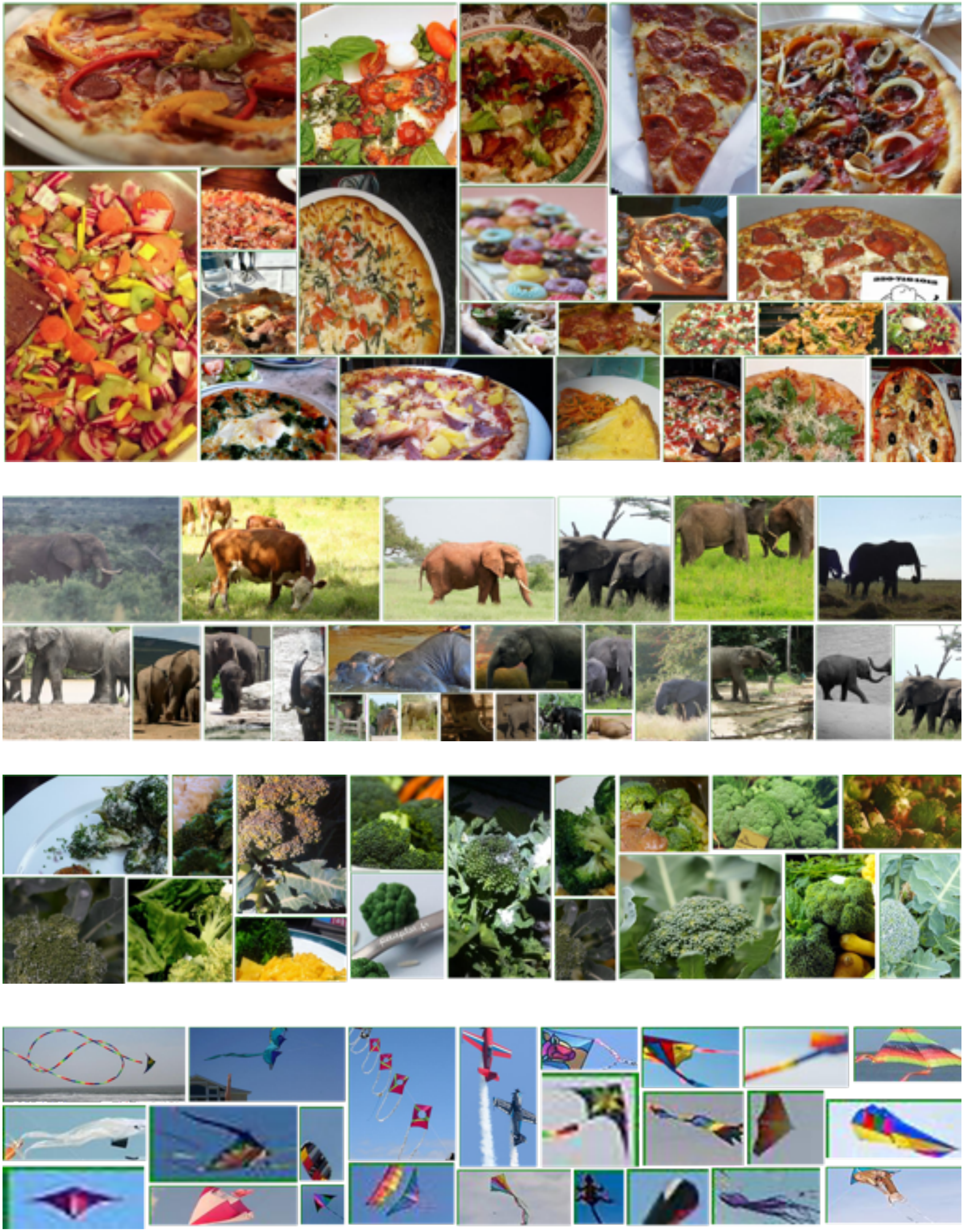

Figure 5: **Qualitative results (cont.)**: Clusters discovered by our approach.

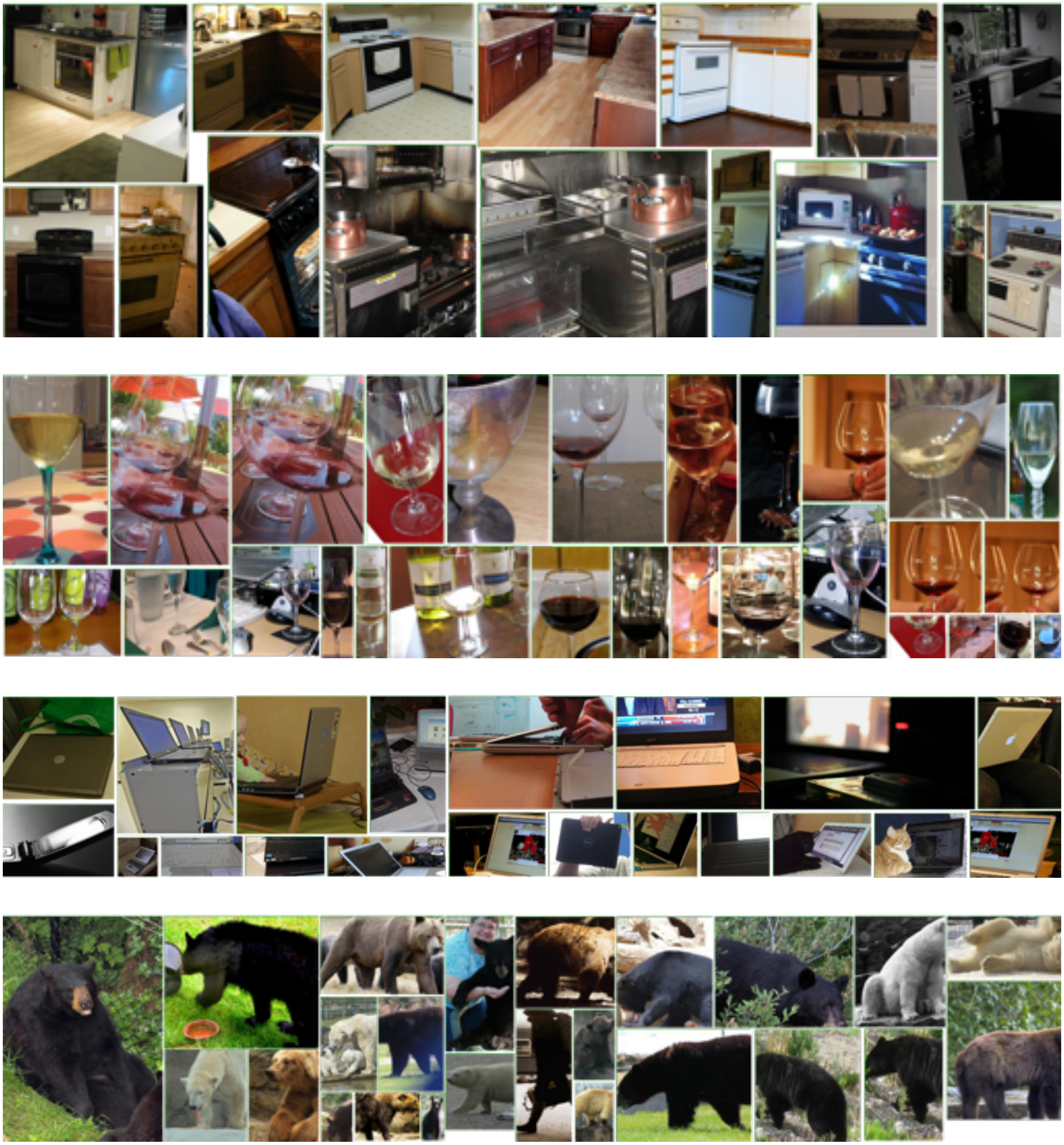

Figure 6: **Qualitative results (cont.)**: Clusters discovered by our approach.

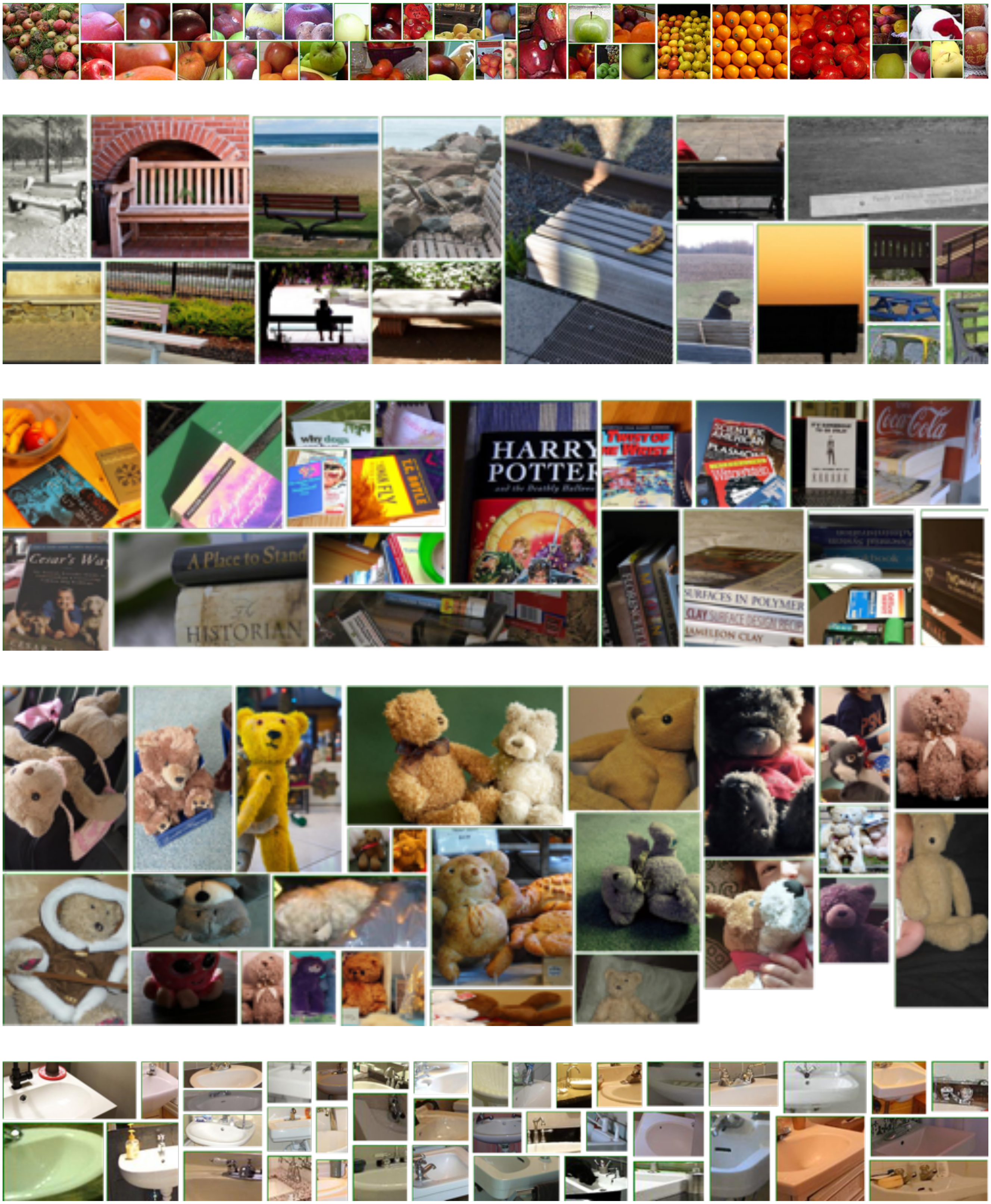

Figure 7: **Qualitative results (cont.)**: Clusters discovered by our approach.

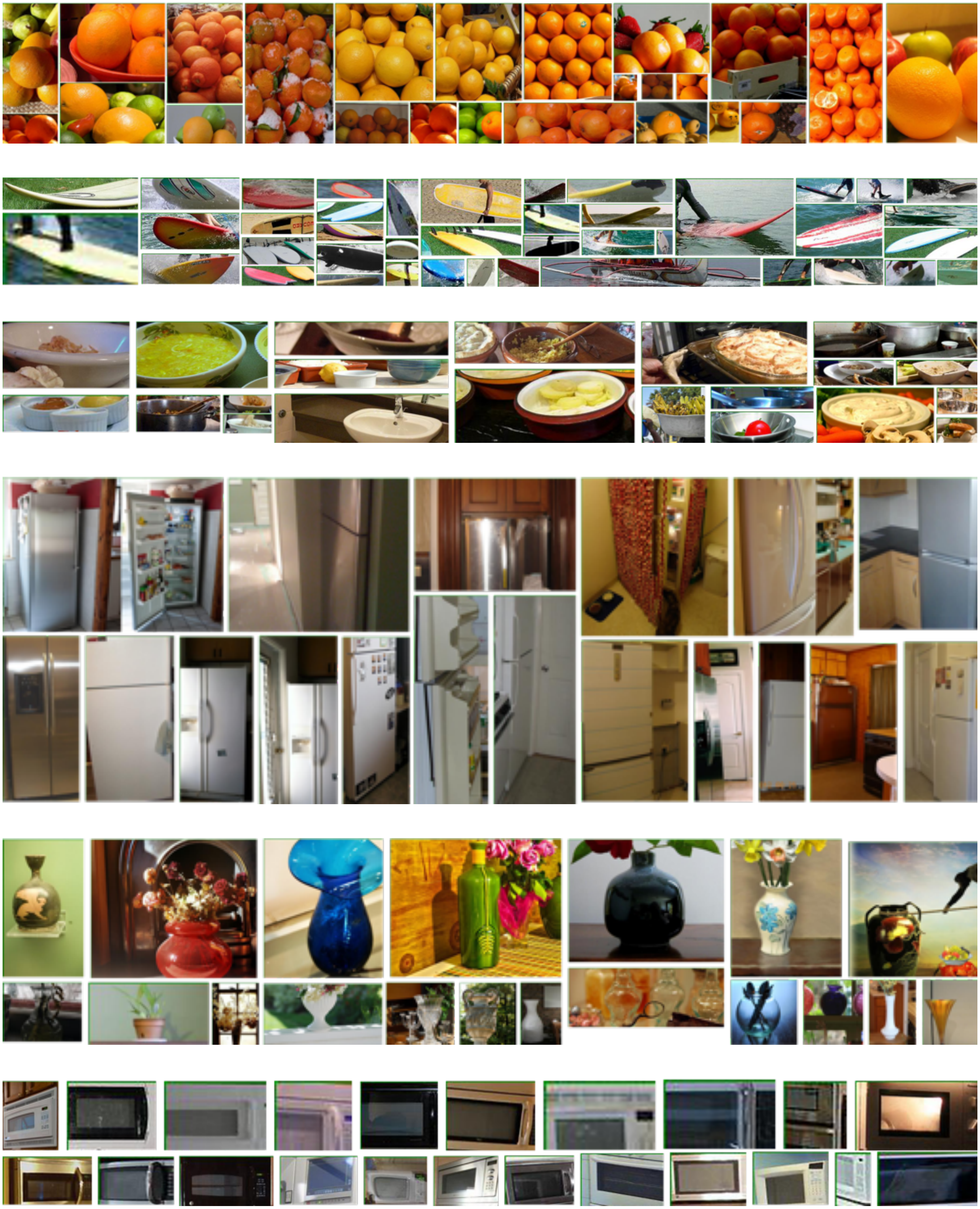

Figure 8: **Qualitative results (cont.)**: Clusters discovered by our approach.

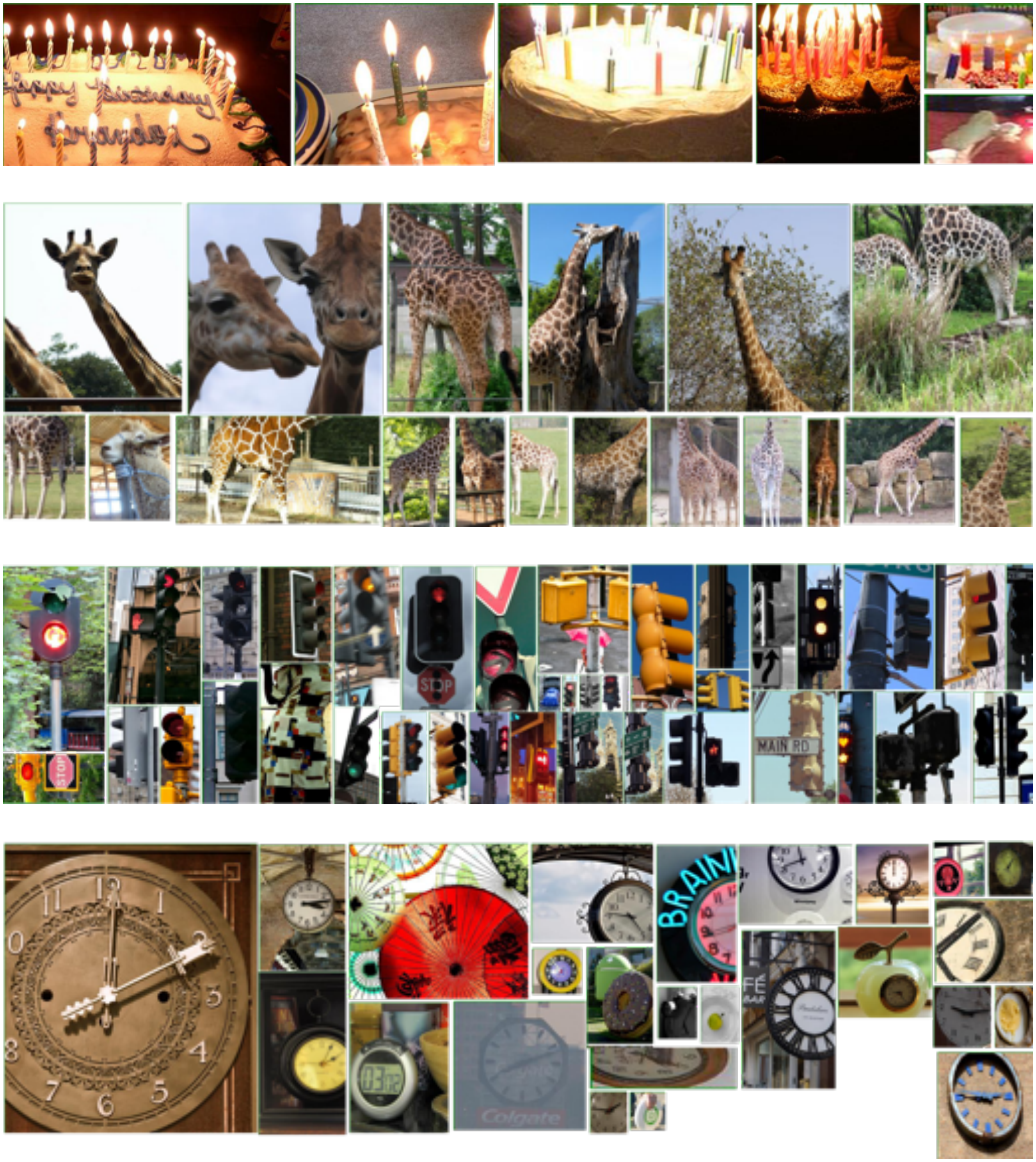

Figure 9: **Qualitative results (cont.)**: Clusters discovered by our approach.

References

[1] Ross Girshick. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1440–1448, Washington, DC, USA, 2015. IEEE Computer Society.

[2] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[3] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[4] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[5] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Deep image category discovery using a transferred similarity function. *CoRR*, abs/1612.01253, 2016.

[6] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[7] R. Arandjelović and A. Zisserman. Object discovery with a copy-pasting GAN. *CoRR*, abs/1905.11369, 2019.

[8] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *CVPR*, 2019.

[9] Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, pages 1721–1728. IEEE, 2011.

[10] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Context as supervisory signal: Discovering objects with predictable context. In *European Conference on Computer Vision (ECCV)*, pages 362–377. Springer, 2014.

[11] Hongwen Kang, Martial Hebert, and Takeo Kanade. Discovering object instances from scenes of daily living. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, page 762–769, USA, 2011. IEEE Computer Society.

[12] Z. Wang and J. Yuan. Simultaneously discovering and localizing common objects in wild images. *IEEE Transactions on Image Processing*, 27(9):4503–4515, Sep. 2018.

[13] C. H. Q. Ding, Xiaofeng He, Hongyuan Zha, Ming Gu, and H. D. Simon. A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 107–114, Nov 2001.

[14] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856, Cambridge, MA, USA, 2001. MIT Press.

[15] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, Aug 2000.

[16] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 370–377 Vol. 1, Oct 2005.

[18] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[19] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[20] E. Belouadah and A. Popescu. Il2m: Class incremental learning with dual memory. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 583–592, Oct 2019.

[21] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[22] Santosh Kumar Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3270–3277, 2014.

[23] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2013.