Semantic Amodal Segmentation

Yan Zhu[1,2], Yuandong Tian[1], Dimitris Metaxas[2], and Piotr Dollár[1]

[1]Facebook AI Research (FAIR)
[2]Department of Computer Science, Rutgers University

Abstract

Common visual recognition tasks such as classification, object detection, and semantic segmentation are rapidly reaching maturity, and given the recent rate of progress, it is not unreasonable to conjecture that techniques for many of these problems will approach human levels of performance in the next few years. In this paper we look to the future: what is the next frontier in visual recognition?

We offer one possible answer to this question. We propose a detailed image annotation that captures information beyond the visible pixels and requires complex reasoning about full scene structure. Specifically, we create an amodal *segmentation of each image: the full extent of each region is marked, not just the visible pixels. Annotators outline and name all salient regions in the image and specify a partial depth order. The result is a rich scene structure, including visible and occluded portions of each region, figure-ground edge information, semantic labels, and object overlap.*

We create two datasets for semantic amodal segmentation. First, we label 500 images in the BSDS dataset with multiple annotators per image, allowing us to study the statistics of human annotations. We show that the proposed full scene annotation is surprisingly consistent between annotators, including for regions and edges. Second, we annotate 5000 images from COCO. This larger dataset allows us to explore a number of algorithmic ideas for amodal segmentation and depth ordering. We introduce novel metrics for these tasks, and along with our strong baselines, define concrete new challenges for the community.

1. Introduction

In recent years, visual recognition tasks such as image classification [22, 16], object detection [10, 35, 13, 33], edge detection [2, 8, 44], and semantic segmentation [36, 30, 26] have witnessed dramatic progress. This has been driven by the availability of large scale image datasets [9, 5, 24] coupled with a renaissance in deep learning techniques with massive model capacity [22, 39, 40, 16]. Given the pace of recent advances, one may conjecture that techniques for many of these tasks will rapidly approach human levels of performance. Indeed, preliminary evidence exists this is already the case for ImageNet classification [20].

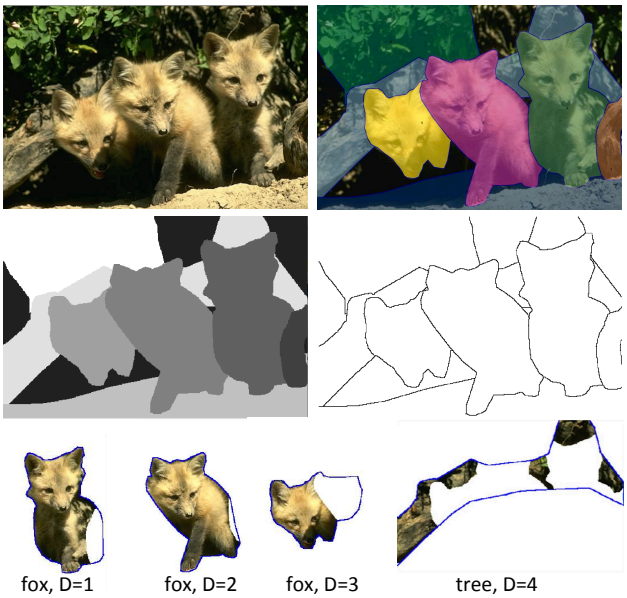


Figure 1: Example of *Semantic Amodal Segmentation*. Given an image (top-left), annotators segment each region (top-right) and specify a partial depth order (middle-left). From this, visible edges can be obtained (middle-right) along with figure-ground assignment for each edge (not shown). All regions are annotated *amodally*: the full extent of each region is marked, not just the visible pixels. Four annotated regions along with their semantic label and depth order are shown (bottom); note that both visible and occluded portions of each region are annotated.

In this work we ask: what are the next set of challenges in visual recognition? What capabilities do we expect future visual recognition systems to possess?

We take our inspiration from the study of the human visual system. A remarkable property of human perception is the ease with which our visual system interpolates information not directly visible in an image [29]. A particularly prominent example of this, and one on which we focus, is *amodal perception*: the phenomenon of perceiving the

whole of a physical structure when only a portion of it is visible [18, 29, 42]. Humans can readily perceive partially occluded objects and guess at their true shape.

To encourage the study of machine vision systems with similar capabilities, we ask human subjects to annotate regions in images *amodally*. Specifically, annotators are asked to mark the full extent of each region, not just the visible pixels. Annotators outline and name all salient regions in the image and specify a partial depth order. The result is a rich scene structure, including visible and occluded portions of each region, figure-ground edge information, semantic labels, and object overlap. See Figure 1.

An astute reader may ask: is amodal segmentation even a well-posed annotation task? More precisely, will multiple annotators agree on the annotation of a given image?

To study these questions, we asked multiple annotators to label all 500 images in the BSDS dataset [2]. We designed the annotation task in a manner that encouraged annotators to consider object relationships and reason about scene geometry. This resulted in agreement between annotators that is surprisingly strong. In particular, our data has higher region and edge consistency than the original BSDS labels. Likewise, annotators tend to agree on the amodal completions. We report a thorough study of human performance on amodal segmentation using this data and also use it to train and evaluate state-of-the-art edge detectors.

In addition to the BSDS data, we annotate a second larger semantic amodal segmentation dataset using 5000 images from COCO [24]. To achieve this scale, each image in COCO was annotated with just one expert annotator plus strict quality control. The dataset is divided into 2500/1250/1250 images for train/val/test, respectively. We introduce novel evaluation metrics for measuring amodal segment quality and pairwise depth-ordering of region segments. We do not currently use the semantic labels for evaluation as they come from an open vocabulary; nevertheless, we show that collecting these labels is key for obtaining high-quality amodal annotations. All train and val annotations along with evaluation code will be publicly released.

Finally, the larger collection of annotations on COCO allows us to train strong baselines for amodal segmentation and depth ordering. To perform amodal segmentation, we extend recent modal segmentation algorithms [31, 32] to the amodal setting. We train two baselines: first, we train a deep net to directly predict amodal masks, second, motivated by [23], we train a model that takes a modal mask and attempts to expand it. Both variants achieve large gains over their modal counterparts, especially under heavy occlusion. We also experiment with deep nets for depth ordering and achieve accuracy over 80%.

Our challenging new dataset, metrics, and strong baselines define concrete new challenges for the community and we hope that they will help spur novel research directions.

Figure 2: *Amodal versus modal segmentation*: The left (red frame) of each image pair shows the modal segmentation of a region (visible pixels only) while the right (green frame) shows the amodal segmentation (visible and interpolated region). In this work we ask annotators to segment regions amodally. Note that the amodal segments have simpler shapes than the modal segments.

1.1. Related Work

Amodal perception [18] has been studied extensively in the psychophysics literature, for a review see [42, 29]. However, amodal completion, along with many of the principles of perceptual grouping, are often demonstrated via simple illustrative examples such as the famous Kanizsa's triangle [18]. To our knowledge, there is no large scale dataset of amodally segmented natural images.

Modal segmentation[1] datasets are more common. The most well known of these is the BSDS dataset [2], which has been used extensively for training and evaluating edge detection [6, 8, 44] and segmentation algorithms [2]. BSDS was later extended with figure-ground edge labels [12]. A drawback of this annotation style is that it lacks clear guidelines, resulting in inconsistencies between annotators.

An alternative to unrestricted modal segmentation is *semantic segmentation* [36, 25, 37], where each image pixel is assigned a unique label from a fixed category set (*e.g.* grass, sky, person). Such datasets have higher consistency than BSDS. However, the label set is typically small, individual objects are not delineated, and the annotations are modal. Notable exception are the StreetScenes dataset [4], which contains a few categories which are labeled amodally, and PASCAL context [28], which uses a large category set.

The closest dataset to ours is the hierarchical scenes dataset from Maire *et al*. [27], which aims to captures occlusion, figure-ground ordering, and object-part relations. The dataset consists of incredibly rich and detailed annotations for 100 images. Our dataset shares some similarities but is easier to collect, allowing us to scale. Likewise, Visual Genome [21] also provides rich annotations, including depth ordering, but does not include segmentation.

Compared to *object detection* datasets [9, 5, 24], our annotation is dense, amodal, and covers both objects and re-

[1] In an abuse of terminology, we use *modal segmentation* to refer to an annotation of only the visible portions of a region. This lets us easily differentiate it from *amodal segmentation* (full region extent annotated).

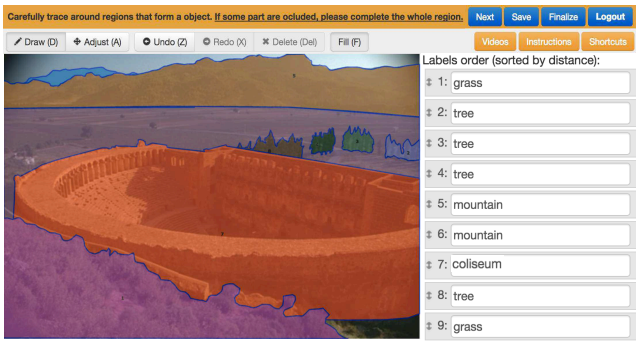


Figure 3: A screenshot of our annotation tool for semantic amodal segmentation (adopted from the Open Surfaces tool [3]).

gions. Related datasets such as Sun [43] have objects annotated modally. LabelMe [34] does have some amodal annotations but not consistently annotated. Only for pedestrian detection [7] are objects often annotated amodally (with both visible and amodal bounding boxes).

We note that our annotation scheme subsumes modal segmentation [2], edge detection [2], and figure-ground edge labeling [12]. As our COCO annotations (5000 images) are an order of magnitude larger than BSDS (500 images) [2], the previous de-facto dataset for these tasks, we expect our data to be quite useful for these classic tasks.

Finally there has been some algorithmic work on amodal completion [14, 15, 38, 19] and depth ordering [41, 45]. Of particular interest, Ke *et al*. [23] recently proposed a general approach for amodal segmentation that serves as the foundation for one of our baselines (see §5). Most existing recognition systems, however, operate on a per-patch or per-window basis, or with a limited receptive field, including for object detection [10, 35, 13], edge detection [6, 8, 44], and semantic segmentation [36, 30, 26]. Our dataset will present challenges to such methods as amodal segmentation requires reasoning about object interactions.

2. Dataset Annotation

For our semantic amodal segmentation, we extend the Open Surfaces annotation tool from Bell *et al*. [3], see Figure 3. The original tool allows for labeling multiple regions in an image by specifying a closed polygon for each; the same tool was also adopted for annotation of COCO [24]. We extend the tool in a number of ways, including for region ordering, naming, and improved editing. For full details, including handling of corner cases, we refer readers to the supplementary. We will open-source the updated tool.

We found four guidelines to be key for obtaining high-quality and consistent annotations: (1) only semantically meaningful regions should be annotated, (2) images should be annotated densely, (3) all regions should be ordered in depth, and (4) shared region boundaries should be marked.

Figure 4: (a) We ask annotators to arrange region depth order. The right panel gives a correct depth order of the two people in the foreground while in the left panel the order is reversed. (b) Shared region edges must be marked to avoid duplicate edges. Unlike regular edges, shared edges do not have a figure-ground side.

These guidelines encouraged annotators to consider object relationships and reason about scene geometry, and have proven to be effective in practice as we show in §4.

(1) Semantic annotation: Annotators are asked to name all annotated regions. Perceptually, the fact that a segment can be named implies that it has a well-defined prototype and corresponds to a semantically meaningful region. This criterion leads to a natural constraint on the granularity of the annotation: material boundaries and object parts (*i.e*. interior edges) should not be annotated if they are not namable. Moreover, under this constraint, annotators are more likely to have a consistent prior on the occluded part of a region. In practice, we found that enforcing region naming led to more consistent and higher-quality amodal annotations.

(2) Dense annotation: Annotators are asked to label an image densely, in particular all foreground object over a minimum size (600 pixels) should be labeled. Of particular importance is that if an annotated region is occluded, the occluder should also be annotated. When all foreground regions are annotated and a depth order specified, the visible and occluded portions of each annotated region are determined, as are the visible and hidden edges.

(3) Depth ordering: Annotators are asked to specify the relative depth order of all regions, see Figure 4a. In particular, for two overlapping regions, the occluder should precede the occludee. In ambiguous cases, the depth order is specified so that edges are correctly 'rendered' (*e.g*., eyes go in front of the face). For non-overlapping regions any depth order is acceptable. Depth ordering encourages annotators to reason about scene geometry, including occlusion, and therefore improves the quality of amodal annotation.

(4) Edge sharing: When one region occludes another, the figure-ground relation is clear, and an edge separating the regions belongs to the foreground region. However, when two regions are adjacent, an edge is shared and has no figure-ground side. We require annotators to explicitly mark shared edges, thus avoiding duplicate edges, see Figure 4b. As with the other criteria, this encourages annotators to reason about object interactions and scene geometry.

We refer readers to the supplementary material for additional details on the annotation tool and pipeline.

	BSDS	COCO
ann/image	5-7	1
regions/ann	7.3	9.2
points/region	64	46
pixel coverage	84%	69%
occlusion rate	62%	61%
occ/region	21%	31%
time/polygon	68s	41s
time/region	2m	2m
time/ann	15m	18m

(a) dataset summary statistics

(b) most common semantic labels

Figure 5: (a) Dataset summary statistics on BSDS and COCO. COCO images are more cluttered, leading to some differences in statistics (*e.g.* higher regions/ann and lower pixel coverage). (b) The top 50 semantic labels in our BSDS annotations. Roughly speaking, the blue words indicate 'things' (person, fish, flower) while the black words indicate 'stuff' (grass, cloud, water).

3. Dataset Statistics

The analysis in this section is primarily based on the 500 images in the BSDS dataset [2], which has been used extensively for edge detection and modal segmentation. Annotating the same images amodally allows us to compare our proposed annotations to the original annotations. While all following analysis is based on these images, we note that the statistics of our annotations on COCO [24] are similar (they differ slightly as COCO images are more cluttered).

Figure 5a summarizes the statistics of our data. Each of the 500 BSDS images was annotated independently by 5 to 7 annotators. On average each image annotation consists of 7.3 labeled regions, and each region polygon consists of 64 points. About 84% of image pixels are covered by at least one region polygon. Of all regions, 62% are partially occluded and average occlusion is 21%.

Annotating a single region takes ∼2 minutes. Of this, half the time is spent on the initial polygon and the rest on naming, depth ordering, and polygon refinement. Annotating an entire image takes ∼15m, although this varies based on image complexity and annotator skill.

Semantic labels: Figure 5b shows the top 50 semantic labels in our data with word size indicating region frequency. The labels give insight into the regions being labeled as well as the granularity of the annotation. Most labels correspond to basic level categories and refer to entire objects (not object parts). Using common terminology [1, 11], we explicitly classify the labels into two categories: 'things' and 'stuff', where a 'thing' is an object with a canonical shape (person, fish, flower) while 'stuff' has a consistent visual appearance but can be of arbitrary spatial extent (grass, cloud, water). Both 'thing' and 'stuff' labels are prevalent in our data (stuff composes about a quarter of our regions).

Shape complexity: One important property of amodal segments is that they tend to have a relatively simple shape compared to modal segments that is independent of scene geometry and occlusion patterns (see Figure 2). We verify this observation with the following two statistics, shape *convexity* and *simplicity*, defined on a segment S:

$$convexity(S) = \frac{Area(S))}{Area(ConvexHull(S))} \quad (1)$$

$$simplicity(S) = \frac{\sqrt{4\pi * Area(S)}}{Perimeter(S)} \quad (2)$$

A segment with a large convexity and simplicity value means it is simple (and both metrics achieve their maximum value of 1.0 for a circle). Table 1 shows that amodal regions are indeed simpler than modal ones, which verifies our hypothesis. Due to their simplicity, amodal regions can actually be more efficient to label than modal regions.

We also compare to the original (modal) BSDS annotations (first column of Table 1). Interestingly, the original BSDS annotations are even simpler than our modal annotations. Qualitatively it appears that the original annotators had a bias for simpler shapes and smoother boundaries.

Edge density: The last row of Table 1 shows that our dataset has fewer visible edges marked than the original BSDS annotation (edge density is the percentage of image pixels that are edge pixels). This is necessarily the case as material boundaries and object parts (i.e. interior edges) are not annotated in our data. Note that in §4 we demonstrate that although our edge maps are slightly less dense, they can be used to effectively train state-of-the-art edge detectors.

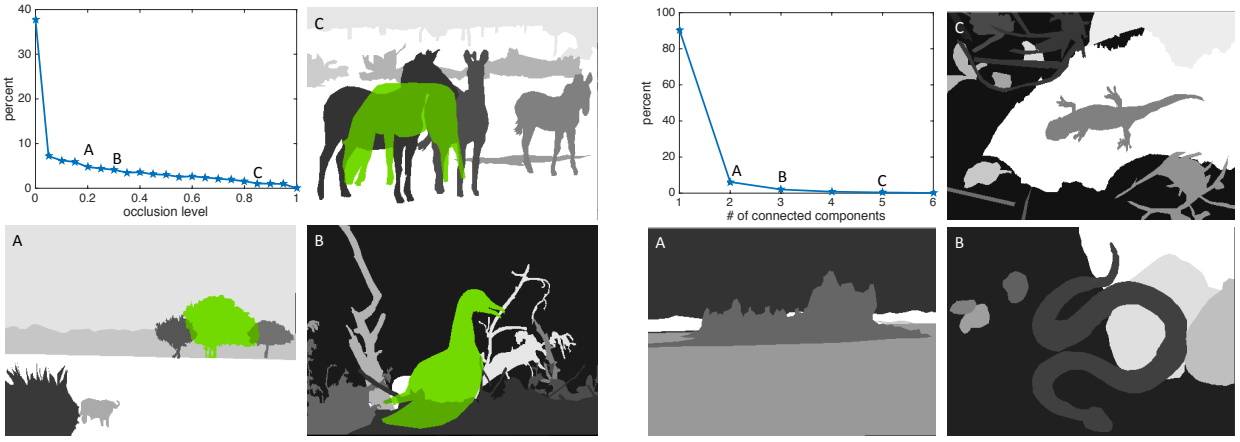
Occlusion: Figure 6a shows a histogram of occlusion level (defined as the fraction of region area that is occluded). Most regions are slightly occluded, while a small portion of regions are heavily occluded. We additionally display 3 occluded examples at different occlusion levels.

Scene complexity: With the help of depth ordering, we can represent regions using a Directed Acyclic Graph (DAG). Specifically, we draw a directed edge from region R_1 to region R_2 if R_1 spatially overlaps R_2 and R_1 precedes R_2 in depth ordering. Given the DAG corresponding to an image annotation, a few quantities can be analyzed.

First, Figure 6b shows the number of connected components (CC) per DAG. Most annotations have only one CC,

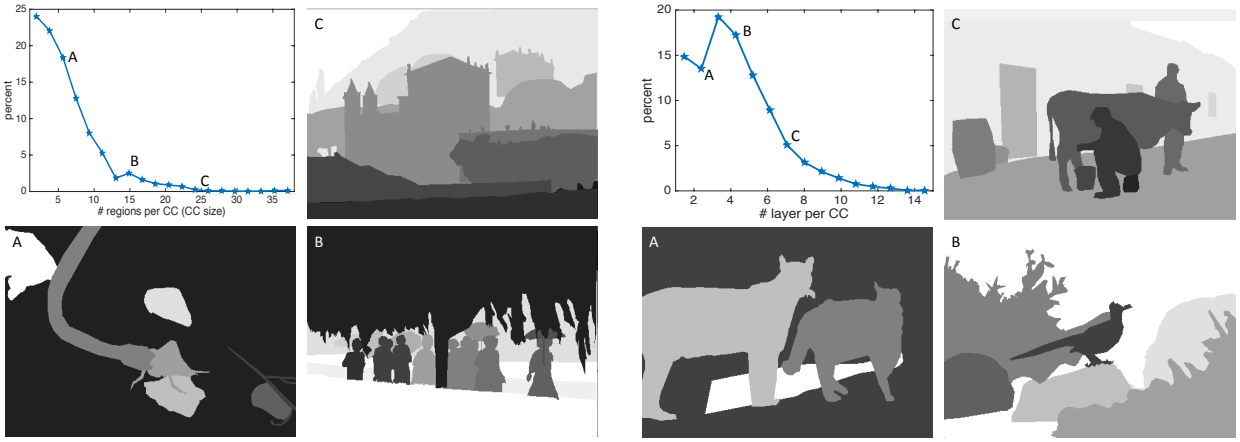
	BSDS			COCO	
	original	modal	amodal	modal	amodal
simplicity	.801	.718	.834	.746	.856
convexity	.664	.616	.643	.658	.685
density	1.80%	1.57%	1.97%	1.71%	2.10%

Table 1: Comparison of shape and edge statistics between modal and amodal segments on BSDS and COCO. Amodal segments tend to have a relatively simpler shape that is independent of scene geometry and occlusion patterns (see also Figure 2). Interestingly, the original BSDS annotations (first column) are even simpler than our modal annotations. Finally the last row reports edge density.



(a) detailed occlusion statistics

(b) number of connected components per annotation



(c) connected components size

(d) number of depth layers per connected component

Figure 6: Detailed dataset statistics. See text for details.

as shown in example A. If regions are scattered and disconnected an image will have more CC's, as in B and C.

The size of a CC measures how many regions are mutually overlapped, which in turns gives an implicit measure of scene complexity. Figure 6c shows a number of examples. More complex scenes (examples B and C) have large CC's.

Finally, the longest directed path of any CC in a DAG characterizes the minimum number of depth layers required to properly order all regions in the DAG. Note that the number of depth layers is often smaller than the size of a CC: e.g. a large CC with numerous non-overlapping foreground objects and a single common background only requires two depth layers. Figure 6d shows the distribution of number of depth layers needed per CC. Most components require only a few depth layers although some are far more complex.

Figure 7 further investigates the correlation between CC size and the minimum number of depth layers necessary to order all regions. We observe that the number of depth layers necessary appears to grow logarithmically with CC size.

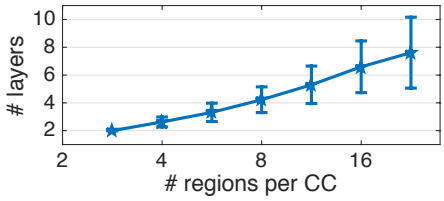


Figure 7: The minimum number of depth layers necessary to represent a connected component (CC). See text for details.

4. Dataset Consistency

We next aim to show that semantic amodal segmentation is a well-posed annotation task. Specifically, we show that agreement between independent annotators is high. Consistency is a key property of any human-labeled dataset as it enables machine vision systems to learn a well defined concept. In the next two sub-sections we analyze our dataset's region and edge consistency on BSDS. As a baseline, we compare to the original (modal) BSDS annotations.

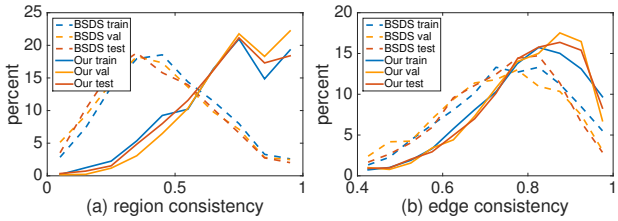


Figure 8: (a) Histogram of pairwise *region consistency* scores for the original *modal* BSDS annotations and our *amodal* regions. (b) Histogram of pairwise *edge consistency* scores for visible edges.

4.1. Region Consistency

To measure region consistency, we use Intersection over Union (IoU) to match regions. The IoU between two segments is the area of their intersection divided by the area of their union. We threshold IoU at 0.5 and use bipartite matching to match two sets of regions. We set each annotation as the ground truth in turn, and for every other annotator we compute precision (P) and recall (R) and summarize the result via the F measure: $F = 2PR/(P+R)$. For n annotators this yields $n(n-1)$ F scores per image.

In Figure 8a we display a histogram of F scores for both the original BSDS *modal* annotations from [2] and the *amodal* annotations in our proposed dataset across each split of the dataset. The region consistency of our amodal regions is substantially higher than the consistency of the original modal regions: median of 0.723 versus 0.425. This is in spite of the fact that our amodal regions include both the visible and occluded portions of each region. We note that the modal region consistency of our annotations is 0.756, slightly higher than for amodal regions, as expected.

A number of factors contribute to the consistency of our regions. Most importantly, we gave more focused instructions to the annotators; specifically, we asked annotators to label only semantically meaningful regions and to label all foreground objects, see §2. Thus there was less inherent ambiguity in the task. Moreover, in modal segmentation, annotation level of detail substantially impacts region agreement.

Figure 9 shows qualitative examples of annotator agreement on individual regions for both visible and occluded portions of a region. Naturally, annotations are most consistent for regions with simple shapes and little occlusions. On the other hand, when the object is highly articulated and/or severely occluded, annotators tend to disagree more.

4.2. Edge Consistency

Given the amodal annotations and depth ordering, along with the constraint that all foreground regions are annotated, we can compute the set of visible image edges. We next verify the quality of the obtained edge maps.

First, to measure edge consistency among annotators, we compute the F score between each pair of annotations, for details see [2]. Figure 8b shows the distribution of the boundary consistency scores. The edges in our amodal dataset are more consistent than edges in the original BSDS annotations (median consistency of 0.795 versus 0.728).

	SE [8]			**HED** [44]		
train / test	ODS	AP	R50	ODS	AP	R50
bsds / bsds	.744	.795	.921	**.787**	.790	.855
ours / bsds	**.747**	**.802**	**.923**	.775	**.793**	**.868**
bsds / ours	.619	.603	.761	.657	**.578**	.697
ours / ours	**.630**	**.630**	**.785**	**.694**	.572	**.752**

Table 2: Cross-dataset performance of two state-of-the-art edge detectors. For SE, training on our dataset improves performance even when testing on the original BSDS edges. For HED, using the same train/test combination maximizes performance. These results indicate that our dataset is valid for edge detection.

While our edges are more consistent, the edges are also less dense (see Table 1). To evaluate the efficacy of using our data for edge detection, we test two popular state-of-the-art edge detectors: structured edges (SE) [8] and the holistically-nested edge detector (HED) [44]. Results for cross-dataset generalization are shown in Table 2. For SE, training on our dataset improves performance even when testing on the original BSDS edges. For HED, using the same train/test combination maximizes performance by a slight margin. These results indicate that our dataset is valid for edge detection. Note, however, that our test set is substantially harder as only semantic boundaries are annotated.

Finally, we measure human performance. As in [2], we take one annotation as the detection and the union of the others as ground truth (note that this differs from the 1-vs-1 methodology used for Figure 8b). On the original BSDS test set, precision/recall/F-Score are .92/.73/.81. Human performance is much higher on our test set, the scores are .98/.83/.90. Of particular interest, however, is the gap between human and machine. On the original BSDS annotations, HED achieves ODS of .79 while human F score is .81, leaving a gap of just .02. On our annotations, however, HED drops to .69 while human F score increases to .90. Thus, unlike the original annotations, our dataset leaves substantial room for improvement of the state-of-the-art.

5. Metrics and Baselines

We aim to develop measures to quantify algorithm performance on our data. We begin by reiterating that our rich annotations subsume many classic grouping tasks, including modal segmentation, edge detection, and figure-ground edge labeling. Indeed, our COCO dataset (5000 images) is an order of magnitude larger than BSDS (500 images), the previous de-facto dataset for these tasks. We encourage researchers to use our data to study these classic tasks; for well-established metrics we refer readers to [2].

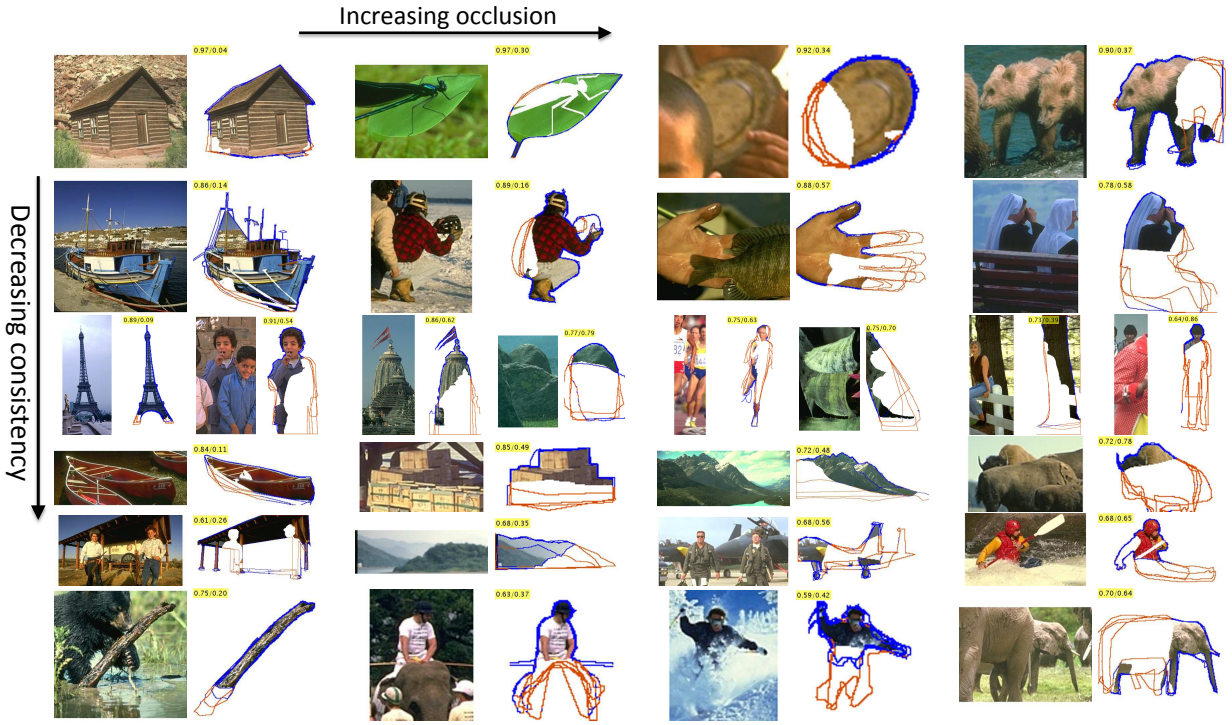


Figure 9: Visualizations of amodal region consistency. The blue edges are the visible edges, while the red edges are the occluded edges. Ground truth is determined by a single randomly chosen annotator. The region consistency score (average IoU score) and the occlusion rate are displayed. Examples are roughly sorted by decreasing consistency vertically and increasing occlusion horizontally.

Here we propose two simple metrics that focus on the most salient aspect of our dataset: the amodal nature of the segmentations. Predicting amodal segments requires understanding object interaction and reasoning about occlusion. Specifically, we propose to evaluate: (1) amodal segment quality and (2) pairwise depth ordering between regions. We additionally define strong baselines for each task.

All experiments are on the 5000 COCO annotations, split into 2500/1250/1250 images for train/val/test, respectively. We evaluate on val and reserve the test images for use in a possible future challenge as is best practice on COCO.

5.1. Amodal Segment Quality

Metrics: To evaluate amodal segments, we adopt a popular metric for object proposals: average recall (AR), proposed in [17] and used in the COCO challenges. To compute AR, segment recall is computed at multiple IoU thresholds (0.5-0.95), then averaged. To extend to our setting, we simply measure the IoU against the *amodal* masks. We measure AR for 1000 segments per image and also separately for things and stuff. Finally, we report AR for varying occlusion levels q: none (q=0), partial ($0<q\leq.25$), and heavy ($q>.25$), comprising 39%, 31% and 30% of the data.

Baselines: We use *DeepMask* [31] and *SharpMask* [32], current state-of-the-art methods for *modal* class-agnostic object segmentation, as our first baselines. Next, inspired by Ke et al. [23] (which is not directly applicable to our setting), we propose a deep network we call *ExpandMask*. ExpandMask takes an image patch and a modal mask generated by SharpMask as input and outputs an amodal mask. Finally, we train a network, which we call *AmodalMask*, to directly predict amodal masks from image patches. ExpandMask and AmodalMask share an identical network architecture with SharpMask (except ExpandMask adds an extra input channel and uses a slightly larger input size). However, while AmodalMask is run convolutionally, ExpandMask is evaluated on top of SharpMask segments.

We use the DeepMask and SharpMask publicly available code and pre-trained models. We implement ExpandMask and AmodalMask on top of the same codebase. Our models are initialized from the SharpMask network trained on the original modal COCO data. We finetune using our amodal training set. We also attempted to finetune our models using synthetic amodal data (*ExpandMask*^{S} and *AmodalMask*^{S}) by randomly overlaying objects masks from the original COCO dataset. For reproducibility, and to elucidate design and network choices, all source code will be released.

Results: AR for all methods is given in Table 3a and qualitative results are shown in Figure 10. SharpMask is a strong baseline, especially for things and under limited occlusion, which is its training setup. With more occlusion, the amodal baselines are superior, indicating these models can predict amodal masks (however, they are worse on unoccluded objects). Using synthetic data improved AR on

	all regions				things only				stuff only			
	AR	AR^N	AR^P	AR^H	AR	AR^N	AR^P	AR^H	AR	AR^N	AR^P	AR^H
DeepMask [31]	.378	.456	.407	.248	.422	.470	.473	.279	.248	.367	.242	.199
SharpMask [32]	.396	**.493**	.428	.242	.448	**.510**	**.501**	.275	.246	.384	.243	.187
$ExpandMask^S$	.384	.460	.415	.256	.427	.474	.480	.284	.258	.374	.250	.212
$AmodalMask^S$	.395	.457	.424	.289	.435	.468	.487	.316	.282	.388	.268	.246
ExpandMask	.417	.480	.428	.327	.456	.495	.488	.351	.305	.387	.278	.289
AmodalMask	**.434**	.470	**.460**	**.364**	**.458**	.479	.498	**.376**	**.366**	**.414**	**.365**	**.346**

(a) amodal segmentation evaluation

	Sharp Mask	Expand Mask	Amodal Mask	Ground Truth	Ground Truth
train-recall	45%	56%	59%	50%	100%
test-recall	41%	51%	54%	100%	100%
area	.696	.703	.719	.715	.715
y-axis	.711	.708	.706	.702	.702
$OrderNet^B$	.753	.764	.770	.770	.765
$OrderNet^M$	.786	.785	.791	.810	.817
$OrderNet^{M+I}$	**.793**	**.802**	**.814**	**.869**	**.883**

(b) depth ordering evaluation

Table 3: (a) Amodal segmentation quality on the COCO validation set for multiple baselines and under no, partial, and heavy occlusion (AR^N, AR^P, AR^H). (b) Accuracy of pairwise depth ordering baselines applied to various segmentations results. See text for details.

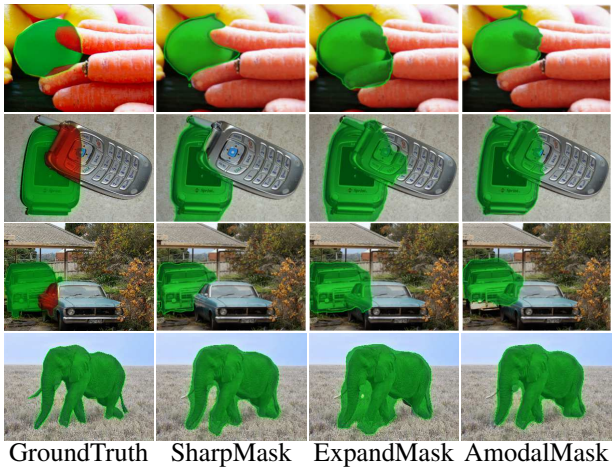


Figure 10: Examples of amodal mask prediction (red indicates occlusion). SharpMask predicts *modal* masks; ExpandMask and AmodalMask predict *amodal* masks. The last row shows an unoccluded object, for which ExpandMask is overzealous.

occluded regions over SharpMask but lagged the accuracy of using real training data. Finally, we note that human accuracy on this task is still substantially higher (see §4).

5.2. Pairwise Depth Ordering

Metrics: Understanding full scene structure is challenging. Instead, we focus on evaluating pairwise depth ordering, which still requires reasoning about object interactions and spatial layout. Specifically, we report the accuracy of predicting which of two overlapping masks is in front. There are 36k/23k overlapping masks in the train/val sets.

Note that we have decoupled depth ordering from mask prediction. Since higher quality masks should be easier to order, we test each ordering algorithm with masks from multiple segmentation approaches. Specifically, for each ground truth mask we first find the best matching mask generated by a segmenter (with IoU of at least 0.5), we then evaluate the depth ordering only on these matched masks.

Baselines: We start with two trivial baselines: order by area (smaller mask in front) and order by y-axis (mask closest to top in back). Next, we implemented a number of deep nets for this binary prediction task: $OrderNet^B$ which takes two bounding boxes as input, $OrderNet^M$ which takes two masks as input, and $OrderNet^{M+I}$ which takes two masks and an image patch. $OrderNet^B$ uses a 3 layer MLP while the other variants use pre-trained ResNet50 models [16] (modified slightly to account for varying number of input channels). We train and test a separate OrderNet model for each set of masks. For each prediction we run inference twice (with input order reversed) and average the results.

Results: We report results in Table 3b. In addition to ordering masks from multiple segmentation algorithms, we also train and test OrderNet on ground truth masks (with varying amount of training data) to capture the role of mask quality and data quantity on ordering accuracy. The naive heuristics (area and y-axis) both achieve about 70% accuracy. OrderNet performs much better, with $OrderNet^{M+I}$ achieving ~80% accuracy on generated masks and ~90% on ground truth. OrderNet benefits from better masks (performance increases in each row moving from left to right), and the percent of recalled pairs also affects results slightly (as there is more data for training). Considering the simplicity of our approach, these results are surprisingly strong.

6. Discussion

We presented a new dataset to study perceptual grouping tasks. The most distinctive feature of our dataset is that regions are annotated amodally: both the visible and occluded portions of regions are marked. The motivation is to encourage amodal perception, and reasoning about object interactions and scene structure. Extensive analysis shows that semantic amodal segmentation is a well-posed annotation task. We also provided evaluation metrics and strong baselines for the proposed tasks. We hope our dataset will help stimulate new research directions for the community.

Acknowledgements

We would like to thank Saining Xie and Yin Li for help with training the HED detector and to Lubomir Bourdev and Manohar Paluri and many others for valuable discussions and feedback.

References

[1] E. H. Adelson and J. R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*. MIT Press, 1991.

[2] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 2011.

[3] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Opensurfaces: A richly annotated catalog of surface appearance. *SIGGRAPH*, 2013.

[4] S. M. Bileschi. *StreetScenes: Towards scene understanding in still images*. PhD thesis, Citeseer, 2006.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] P. Dollár, Z. Tu, and S. Belongie. Supervised learning of edges and object boundaries. In *CVPR*, 2006.

[7] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 2011.

[8] P. Dollár and C. L. Zitnick. Fast edge detection using structured forests. *PAMI*, 2015.

[9] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 2010.

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[11] D. A. Forsyth, J. Malik, M. M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. *Finding pictures of objects in large collections of images*. Springer, 1996.

[12] C. Fowlkes, D. Martin, and J. Malik. Local figure–ground cues are valid for natural images. *Journal of Vision*, 2007.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[14] R. Guo and D. Hoiem. Beyond the line of sight: labeling the underlying surfaces. In *ECCV*, 2012.

[15] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In *CVPR*, 2013.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[17] J. Hosang, R. Benenson, P. Dollár, and B. Schiele. What makes for effective detection proposals? *PAMI*, 2015.

[18] G. Kanizsa. *Organization in vision: Essays on Gestalt perception*. Praeger Publishers, 1979.

[19] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Amodal completion and size constancy in natural scenes. In *ICCV*, 2015.

[20] A. Karpathy, 2015. http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/.

[21] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[22] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural nets. In *NIPS*, 2012.

[23] K. Li and J. Malik. Amodal instance segmentation. In *ECCV*, 2016.

[24] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common objects in context. *PAMI*, 2015.

[25] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 2011.

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[27] M. Maire, S. X. Yu, and P. Perona. Hierarchical scene annotation. In *BMVC*, 2013.

[28] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segm. in the wild. In *CVPR*, 2014.

[29] S. E. Palmer. *Vision science: Photons to phenomenology*. MIT press Cambridge, MA, 1999.

[30] P. O. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *ICML*, 2014.

[31] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In *NIPS*, 2015.

[32] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *ECCV*, 2016.

[33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[34] B. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: A database and web-based tool for image annotation. *IJCV*, 2008.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[36] J. Shotton, J. Winn, C. Rother, and A. Criminisi. *TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segm. In *ECCV*, 2006.

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[38] N. Silberman, L. Shapira, R. Gal, and P. Kohli. A contour completion model for augmenting surface reconstructions. In *ECCV*, 2014.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[40] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[41] J. Tighe, M. Niethammer, and S. Lazebnik. Scene parsing with object instances and occlusion ordering. In *CVPR*, 2014.

[42] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt. A century of Gestalt psychology in visual perception. *Psychological Bulletin*, 2012.

[43] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[44] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[45] Y. Yang, S. Hallman, D. Ramanan, and C. Fowlkes. Layered object detection for multi-class segm. In *CVPR*, 2010.