

Fluid Annotation: A Human-Machine Collaboration Interface for Full Image Annotation

Mykhaylo Andriluka* Jasper R. R. Uijlings* Vittorio Ferrari

Google Research
Zürich, Switzerland

ABSTRACT

We introduce Fluid Annotation, an intuitive human-machine collaboration interface for annotating the class label and outline of every object and background region in an image¹. Fluid annotation is based on three principles: *(I) Strong Machine-Learning aid*. We start from the output of a strong neural network model, which the annotator can edit by correcting the labels of existing regions, adding new regions to cover missing objects, and removing incorrect regions. The edit operations are also assisted by the model. *(II) Full image annotation in a single pass*. As opposed to performing a series of small annotation tasks in isolation [51, 68], we propose a unified interface for full image annotation in a single pass. *(III) Empower the annotator*. We empower the annotator to choose what to annotate and in which order. This enables concentrating on what the machine does not already know, i.e. putting human effort only on the errors it made. This helps using the annotation budget effectively. Through extensive experiments on the COCO+Stuff dataset [11, 51], we demonstrate that Fluid Annotation leads to accurate annotations very efficiently, taking 3× less annotation time than the popular LabelMe interface [70].

KEYWORDS

Computer vision, image annotation, human-machine collaboration

1 INTRODUCTION

The need for large amounts of high-quality training data is quickly becoming a major bottleneck in deep learning. Popular computer vision models continue to grow in size (e.g. [30, 31, 46, 66, 76, 79]) and larger amounts of data continues to improve accuracy [11, 33, 54, 72, 78]). Annotation is especially expensive for models requiring training images annotated with the class label and outline of every object and background region [15, 52, 53, 87]. For example, annotating one image of the COCO dataset [51] required 80 seconds per object to draw a polygon on it [51], plus 3 minutes annotating background regions [11], for a total of 19 minutes on average. Similarly, fully annotating one image of the Cityscapes dataset [17] took 1.5 hours.

In this paper we propose Fluid Annotation, a new human-machine collaboration interface for annotating all objects and background regions in an image. The goal is to have a very efficient and natural interface, which can produce high quality annotations with much less human effort than traditional manual interfaces (Fig. 1).

Fluid Annotation is based on three design principles:

¹Live demo of the interface is available at fluidann.appspot.com

*The authors contributed equally.

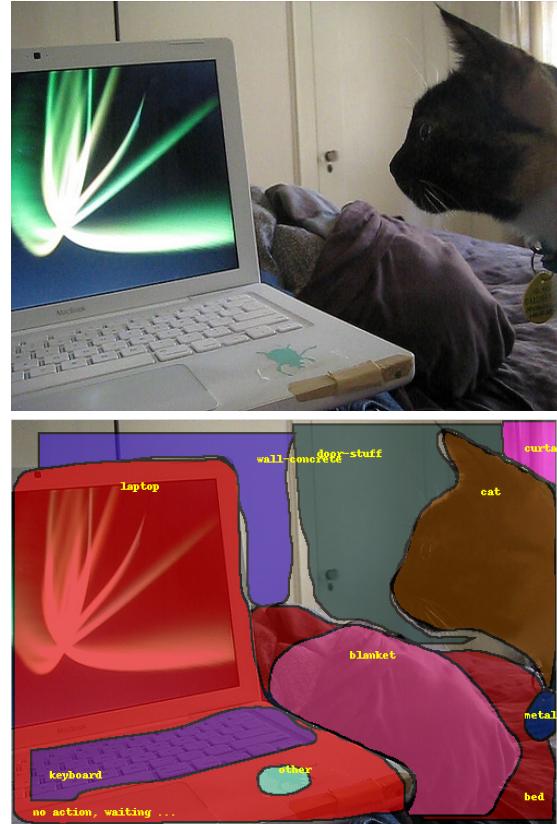


Figure 1: Example of an annotation result: original image (top) and full annotation of objects and background obtained with just 9 annotation actions using our approach (bottom).

(I) Strong Machine-Learning aid. Popular semantic segmentation datasets [11, 17, 22, 51, 58, 92] are annotated fully manually which is very costly. Instead Fluid Annotation starts from the output of a neural network model [30], which the annotator can edit by correcting the label of existing regions, adding new regions to cover missing objects, and removing incorrect regions (Fig. 2). By starting from the most likely machine-generated output, and supporting quick and intuitive editing operations, Fluid Annotation enables full image annotation in a short amount of time.

(II) Unified interface for full image annotation in a single pass. Many datasets are annotated using a series of micro-tasks such

as indicating object presence in an image [19, 51, 68], clicking on instances of a specific class [51], or drawing a polygon or a box around a single instance [22, 51, 77]. Correspondingly, previous ML-aided interfaces focus on a single micro-task, such as segmenting individual objects [2, 9, 13, 34, 35, 49, 59, 67, 90] or annotating bounding boxes [62], or they focus on selecting which micro-task to assign to the annotator [43, 69, 83]. In contrast, with Fluid Annotation we propose a single, unified ML-aided interface to do full image annotation in a single pass.

(III) Empower the annotator. In most annotation approaches there is a fixed sequence of annotation actions [11, 17, 22, 51, 58, 92] or the sequence is determined by the machine [43, 69, 83]. In contrast, Fluid Annotation empowers the annotator: he sees at a glance the best available machine segmentation of all scene elements, and then decides what to annotate and in which order. This enables to focus on what the machine *does not already know*, i.e. putting human effort only on the errors it made, and typically addressing the biggest errors first. This helps using the annotation budget effectively , and also steers towards labeling hard examples first. Focusing on hard examples is known to be beneficial to improve the model later on (e.g. [23, 25, 74]).

Our contributions are: (1) We introduce Fluid Annotation, an intuitive human-machine collaboration interface for fully annotating an image in a single pass. (2) By using simulated annotators, we demonstrate the validity of our approach and optimize the effectiveness of our interface. (3) Using expert human annotators, we compare our Fluid Annotation interface with the popular LabelMe interface [70] and demonstrate that we can produce annotations of similar quality while reducing time by a factor of 3×.

2 RELATED WORK

Weak supervision. A common approach to reduce annotation effort is to use weakly labeled data. For example, several works train object class detectors from image-level labels only (i.e. without annotated bounding boxes) [7, 16, 20, 29, 38, 93]. Other works require clicking on a single point per object in images [61], or per action in video [56]. Semantic segmentation models have been trained using image-level labels only [42, 64], using point-clicks [4, 5, 86], from boxes [40, 55, 60] and from scribbles [50, 89].

A recent variant of weakly supervised learning is the so-called "webly supervised learning", where one learns from large amounts of noisy data crawled from the web [6, 24, 36, 47, 48]. While large amounts of images with image-level labels can be obtained in this manner, full-image segmentation annotations cannot be readily crawled from the web.

Human-machine collaborative annotation. Several works have explored interactive annotation, where the human annotator and the machine model collaborate. In weakly supervised works the human provides annotations only once before the machine starts processing. Interactive annotation systems instead iterate between humans providing annotations and the machine refining its output.

Most works on interactive annotation focus on a single, very specific task. In particular, many works address segmenting a single object instance by combining a machine model and user input within an interactive framework [2, 9, 13, 34, 35, 49, 59, 67, 90]. Typically, the machine first predicts an initial segmentation, which

is then corrected by clicks [34, 49, 90], scribbles [9, 59, 67], or by editing polygon vertices [2, 13]. The machine then updates the segmentation based on the user input and the process iterates until the user is satisfied. Other works address other specific tasks, such as annotating bounding boxes of a given class known to be present in the image [62], and fine-grained image classification through attributes [8, 10, 63, 85]. Instead of focusing on a specific task, we propose a full image annotation interface, covering the class label and outlines of all objects and background regions.

Another research direction focuses on selecting which micro-task to assign to the annotator [43, 69, 83]. In [43] they train an agent to automatically choose between asking an annotator to manually draw a bounding box or to verify a machine-generated box. In [69] the set of micro-tasks also includes asking for an image-level label and finding other missing instances of a class within the same image.

Active learning. Active learning systems start with a partially labeled dataset, train an initial model, and ask human annotations for examples which are expected to improve the model accuracy the most. Active learning has been used to train whole-image classifiers [37, 39, 44, 65], object class detectors [84, 91], and semantic segmentation [75, 82]. While active learning focuses on which examples to annotate, we explore creating a human-machine collaboration interface for full image annotation.

3 FLUID ANNOTATION

3.1 The task: full image annotation

We address the task of full image annotation [11, 17, 41, 57]: We want to annotate every object and background region in the image. The set of classes to be annotated is fixed and predefined. We consider both *thing* classes, countable objects with a well-defined shape (e.g. *cat* and *bus*), and *stuff* classes, which are amorphous and have no distinct parts (e.g. *grass* and *road*). For thing classes, each individual object instance needs to be annotated by its class label and a region defining its spatial extent. For example, there might be 3 cats in the image, each annotated as its own separate region. For stuff classes, all pixels need to be annotated with their class label, but there is no concept of instances. For example, all grass pixels in an image need to be labelled as grass, but it does not matter whether they are annotated as one single big region or split into multiple regions. An example of a fully annotated image is illustrated in Fig. 1.

This task definition corresponds exactly to the *Panoptic Segmentation* task set out by [41]. It subsumes most previous tasks in image understanding, including image classification (only image-level labels, no localization) [27], object detection (only bounding-boxes, no outlines) [22, 45, 68], instance segmentation (only things, no stuff) [26, 28], semantic segmentation (no separation between different object instances of the same class) [73, 80].

3.2 Overview of the Fluid Annotation interface

Fig. 2 gives an overview of the fluid annotation interface and the user interactions it supports. Given an image to be annotated (Fig. 2a), we first apply a neural network model [30] to produce a large, overcomplete set of overlapping segments, aiming at covering most objects and stuff regions in the image. We call this the *proposal*



Figure 2: Example of the annotation process starting from the automatic initialization (top left) and progressing towards the final result (bottom right). Yellow circle marks the location of the mouse click.

set. We then create the most likely machine-generated annotation by automatically choosing a subset of the proposal set, which we call the *active set*. We resolve pixel-level ambiguities by introducing a depth ordering on the active set such that each pixel is only attributed to a single segment. The active set with its depth ordering defines the initial annotation. This is presented to the user through the Fluid Annotation interface, which is deliberately kept simple and shows only the image with the annotation overlaid, as shown in Fig. 2.

The annotator can edit the current annotation by carrying out a series of actions out of the following set: (a) change the label of an active segment, (b) change the depth order of an active segment, (c) remove an active segment, (d) add a segment to the active set, by selecting one out of the proposal set. The annotator is free to choose which actions to perform and in which order to perform them. The resulting interface enables efficient, full image annotation: only 9 actions are needed for the example in Fig. 2.

3.3 Interface elements: segments and labels

Fluid Annotation operates on a proposal set of segments with their corresponding labels. For Fluid Annotation to work well, we have two requirements: (I) the proposal set should cover most of the objects and stuff regions in the image. This requirement can be satisfied by operating on a large and diverse set of segments [3, 12, 21, 81]. (II) Each segment in the proposal set should come with a corresponding *class label* and *segment score*, where the segment score represents the confidence of the class label. This

requirement can be satisfied by using one of the strong modern instance segmentation models [14, 18, 30].

In practice we create our proposal set using Mask-RCNN [30] with Inception-ResNet [79] using the TensorFlow implementation of [32]. Mask-RCNN originally covered only thing classes. We train a second Mask-RCNN model dedicated to stuff classes by treating connected stuff regions as positive training examples (i.e. two disjoint “grass” regions in the same image will result in two separate training examples). To increase the number of output segments during inference we make two small modifications: (1) increase the number of proposals coming from the RPN from 300 to 500; (2) for the final masks, we keep the per-class Non-Maximum-Suppression of the final boxes to $IoU > 0.6$, but increase the total number output masks from 100 to 500. As we use two models this leads to a proposal set of size 1000 per image.

3.4 Interface actions

We now specify in more detail the editing actions available to the annotator (Fig. 2).

Add segment. This action is initiated by left-clicking anywhere in the image. We defined as *valid* all segments in the proposal set that contain the clicked point. To reduce the number of redundant segments we may sort the valid segments by detection score while ignoring their class label, and then perform standard non-maximum suppression (NMS). Note that segments suppressed this way for a given add action may still be added to the annotation by clicking on another location.

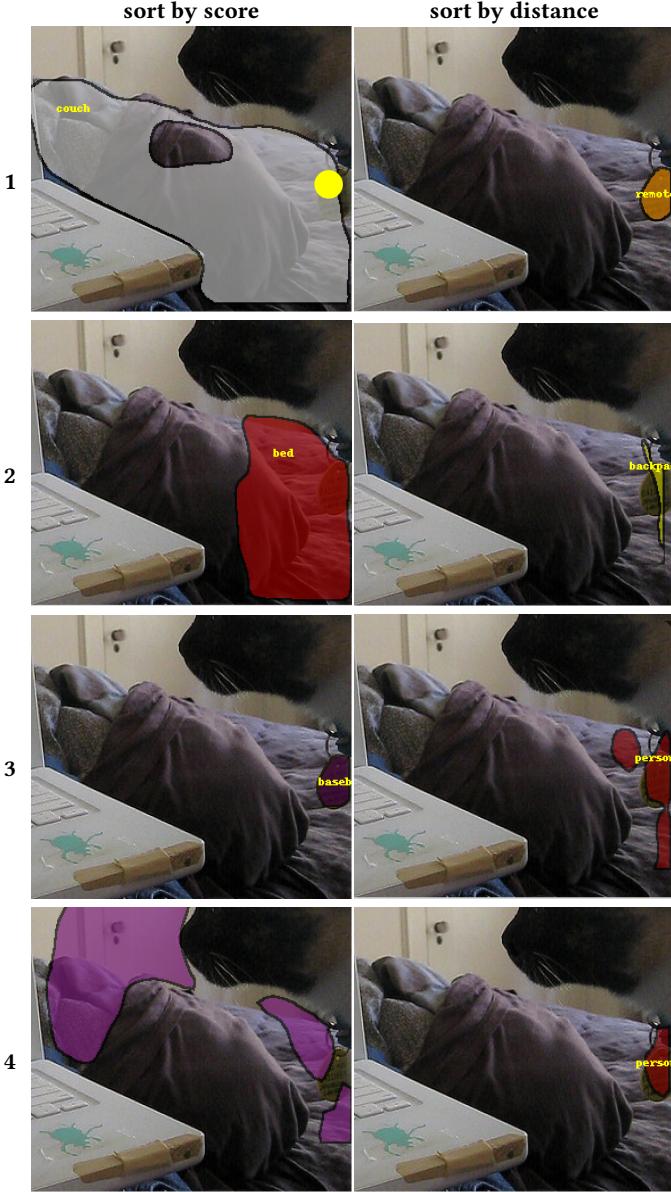


Figure 3: Sequence of segments displayed to the annotator in the *sort-by-score* setting (top row) and *sort-by-distance* setting (bottom row). Mouse click position is marked with the yellow circle in the image in the top-left.

We show one valid segment to the user overlaid on the current annotation. The user can scroll through the valid segments with the mouse wheel (Fig. 3). We experiment with two possible orderings for scrolling: order by segment score or order by the Mahalanobis distance between the click location and the segment’s center of mass. Since the Mahalonobis distance is defined by the spatial variance of the points belonging to the segment, the resulting ordering is affected by both its location and shape. To confirm the selection of the currently visible segment, the user left-clicks a second time.

This adds the selected segment to the active set and moves it in front of all other segments.

We evaluate different variants of the "Add" action in Sec. 5.1, and show that both NMS and distance-based ordering improve the efficiency of the annotation process.

Remove segment. The annotator can remove any active segment by right-clicking on it.

Change label. The annotator can change the label of an active segment by hovering the mouse over it and pressing any key on the keyboard. A drop-down menu appears from which the annotator can choose a label. Instead of showing the complete label set (171 labels in our experiments), we build a shortlist of likely labels for this segment on-the-fly. More precisely, we consider the labels of all segments that contain the current mouse position, sorted by their score. If the correct label is not in the shortlist, the annotator can enter it manually.

Change depth order. The active set of segments is ordered by depth so that each pixel is only attributed to a single segment. The annotator can change the depth ordering of a segment by hovering the mouse over it and scrolling the mouse wheel. Changing depth order is useful as we operate on a large set of overlapping segments, which may be in the wrong depth order. A good example is Fig. 2b, where re-ordering the "couch" to be behind the "person" improves the annotation.

Hide annotations. Sometimes it is hard to see the small details of the image in the interface with the segments overlaid. Pressing the space-bar temporarily hides all annotations.

3.5 Initialization

Fluid Annotation starts from a machine-generated initialization (Fig. 2b), which we construct as follows. We start with an active set containing only the single segment with the highest score. Then, we add the next highest-scored segment, provided that any of its pixels are not yet covered by segments already in the active set. We repeat this process until all segments have been considered. The depth ordering of the segments correspond to the order in which they have been added to the active set.

In Sec. 5.1 we experimentally test whether starting Fluid Annotation from this automatic initialization is beneficial compared to starting from scratch (i.e. an empty active set).

4 SIMULATOR

To efficiently explore various design options in our system we create a simulation environment that aims to imitate the human annotation process. For realism, the simulator operates on the very same Fluid Annotation system as the human would, including the same editing actions (Sec. 3.4).

In order to mimic an annotator the simulator has access to the ground-truth full image annotation which enables evaluating candidate actions with respect to a measure of annotation quality. We use the panoptic quality metric [41] as a measure for annotation quality in our simulation. The panoptic metric is particularly well suited for our purposes since it jointly optimizes the precision, recall and pixel-level accuracy. We refer to [41] for further details on the panoptic metric.

Choosing between edit actions. Our simulator uses a greedy strategy to choose between edit actions for optimizing the annotation quality. Before choosing an action, we generate a pool of candidate actions. For each segment in the active set, we generate 3 candidate action: (1) its “remove” action; (2) one “change depth order” action by choosing the closest reordering which improves annotation quality; (3) one “change label” action by setting the segment label to the label of the best matching ground-truth segment. In addition, for each ground-truth segment which does not have a matching segment in the active set, we generate an “add segment” candidate action. We do this by first simulating the mouse click and then scrolling through the set of segments available at the click location (valid segments). We stop scrolling as soon as the annotation quality improves. Out of the pool of candidate actions, we execute the one that leads to the largest improvement in annotation quality.

Mouse-position simulation. For each edit action the simulator must generate image coordinates for the mouse cursor. Note that an edit action either targets one of the segments in the *active set* (i.e. “remove”, “change depth order”, and “change label”), or targets one of the ground-truth segments (i.e. “add segment”). To simulate positioning the mouse over a target segment we sample its position from a Gaussian distribution defined by the locations of the pixels of the target segment. If the sampled location is not on the target segment, we simply re-sample.

5 EXPERIMENTAL RESULTS

In this section we evaluate our annotation approach. We first start with simulation experiments (Sec. 5.1). Employing a simulator enables us to efficiently explore a broad range of possible settings for the interface. In the second batch of experiments we evaluate the performance of Fluid Annotation when operated by human annotators, in the best settings found during the simulations (Sec. 5.2). **Evaluation metrics.** To perform one edit action in our system the annotator has to perform several interactions with the GUI. We denote these as *micro-actions*. For example, “remove” amounts to a single micro-action corresponding to a mouse click on a segment, whereas the “add” action is composed of a click on a new location, several scrolls of the mouse wheel to sweep through the candidate segments, and another click to confirm the selection of the current candidate. To evaluate the effectiveness of the annotation process we measure the quality of the annotation as a function of the number of micro-actions spent to reach a that level of quality (averaged over all images). We measure annotation quality by recall at $\text{IoU} > 0.5$. For thing classes, each object instance contributes separately to recall. For each stuff class we measure IoU by treating all pixels of that class in an image as a single region (both in the ground-truth and in the output of our method). This matches the panoptic metric [41]. For the human annotation experiments, we also include time measurements in seconds, and we measure agreement between multiple annotators by the average pixel accuracy (as done by [11, 17, 92]).

Dataset. We evaluate our interface on the COCO 2017 validation set (5K images). We use the ground-truth annotation provided by [11, 51] that includes 80 thing classes [51] and 91 stuff classes [11]. This data is highly complete: 94% of all pixels are annotated in the ground-truth [11].

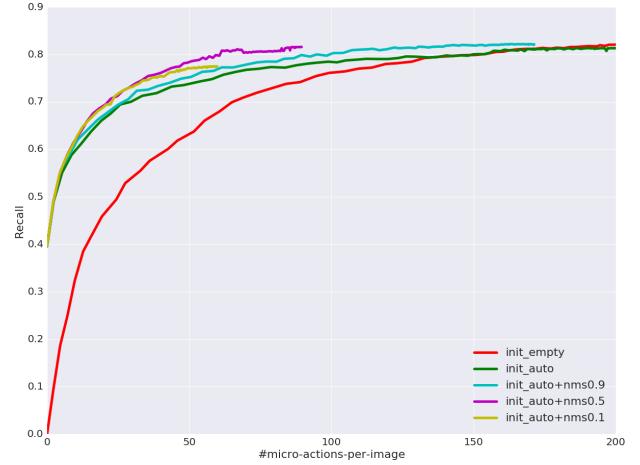


Figure 4: Performance using the basic settings of our system with and without automatic initialization, and comparison for different NMS thresholds.

We randomly split the validation set into 500 images that we use to explore various settings of the interface, and a hold-out set of 4500 images on which we perform the final evaluation of the best performing settings. To evaluate performance with human annotators we use smaller sets of 20 and 25 randomly sampled images from our hold-out set.

We train our segmentation model (Mask-RCNN) on the COCO 2017 challenge training sets [11, 51]. We train one model on the object detection challenge training set (120k images with 80 thing classes), and a second model on the stuff challenge training set (40k images with 91 stuff classes). These training sets do not overlap with the validation set.

5.1 Results using simulations

We first evaluate our Fluid Annotation interface using its basic settings: for the “add” action we do not apply NMS and we order the segments by their score (Sec. 3.4). We consider both a machine-generated initialization (*init-auto*) and starting from scratch (*init-empty*) (Sec. 3.5). Intuitively, a good initialization should save annotation time. A bad initialization would require many corrections and it may increase time instead.

Results are shown in Fig. 4. First of all, we observe that recall rapidly increases during the first few micro-actions: using only 50 micro-actions one reaches 63% (*init-empty*) or 74% (*init-auto*) recall. This demonstrates that, even in its basic settings, our interface is very effective especially in the beginning of the annotation process.

Second, the machine-generated initialization kickstarts the annotation process already at 40% recall, and this is more than doubled to 83% after 200 micro-actions. Example initialization mistakes are shown in Fig. 2 (b): The “door” is mistaken as part of the “wall”, and the “blanket” is labeled as “person”, while the “keyboard” and “metal” medal of the cat were missed altogether. This demonstrates that our machine segmentation model has still plenty to learn and can benefit from more training data.



Figure 5: Distribution of actions and micro-actions.

Finally, *init-auto* leads to a substantially better recall curve compared to *init-empty*: after 50 micro-actions it results in 74% recall whereas *init-empty* only leads to 63% recall. This shows that the machine-generated initialization is clearly beneficial and we adopt this in all subsequent experiments.

Statistics of actions and micro-actions. To identify directions for improving over the basic settings we examine the distributions of actions and micro-actions performed during annotation (always using *init-auto*). In Fig. 5 we show (a) the proportions of different action types performed, (b) the distribution of micro-actions load per action type, and (c) the distribution of micro-actions for action "Add". Note how the lion share of micro-actions is consumed by the "Add" action, and that most of these micro-actions are mouse wheel "scroll" micro-actions used to select the best-fitting segment at the clicked location. The next experiments therefore focus on this segment selection process.

Reducing redundancy with NMS. We apply NMS on the valid set of segments to reduce redundancy (Sec. 3.4). Results for various NMS thresholds are shown in Fig. 4. We see that for a small threshold of 0.1 too many segments are removed, hurting the overall recall. On the other extreme, a high threshold of 0.9 does preserve recall, but removes only a few segments. A moderate NMS threshold of 0.5 instead allows to preserve recall while substantially reducing the number of segments that need to be considered during the "Add" action. At this threshold, the top recall is reached after just 61 micro-actions, compared to more than 150 micro-actions without NMS. We therefore adopt *init-auto+nms0.5* in all subsequent experiments.

Sorting by score vs by distance. In Fig. 6 we compare the two segment orderings introduced in Sec. 3.4 (i.e. by segment score, or by the Mahalanobis distance between the click location and the segment's

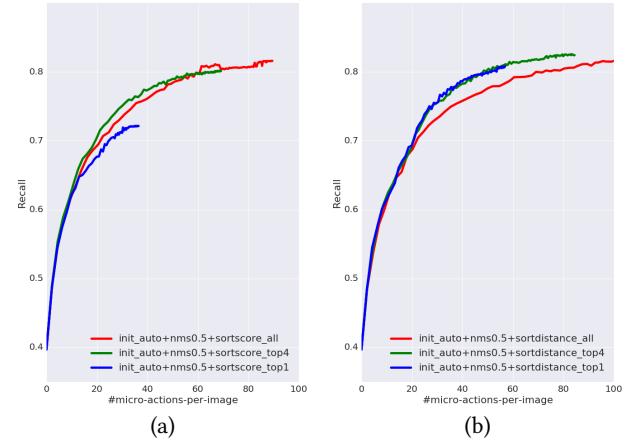


Figure 6: Ordering of segments according to Mask-RCNN score (a) vs. Mahalanobis distance between the click location and the segment center (b).

center). To highlight the ability of each ordering to place the correct segment early we limit the maximum number of segments available to annotator during the "Add" action. We denote our system settings with limit to N top segments and ordering by score as *init-auto+nms0.5+sortscore-topN*, and settings with ordering by distance as *init-auto+nms0.5+sortdistance-topN*.

Without top N limiting the curves for both orderings are about the same and achieve 82% recall in 90 micro-actions (compare the red curves in Fig. 6 (a) and (b)). When setting the limit to top 4, distance ordering achieves a higher final recall than score ordering, 83% vs 80% (compare green curves). It is also more efficient: at 40 micro-actions, it yields 79% recall instead of 77%. We conclude that distance ordering and using top 4 is the most efficient settings, and it does not reduce final recall. We use this for subsequent experiments.

To better understand this result, consider the extreme case of limiting to the top 1 segment only. When the annotator clicks near the center of a missing object, the top 1 segment ordered by distance will be the right one (assuming it is in the proposal set). In contrast, when ordering by score, the top 1 segment is potentially unrelated to the missing object. In the worst case, the overall highest scored segment might occupy the whole image. In that case, it will be the only available segment, regardless of where the annotator clicks.

Validation of results on hold-out set. We verify the effectiveness of the chosen settings on the hold-out set of 4500 images. In particular, Fig. 7 compares *init-auto*, *init-auto+nms0.5*, and *init-auto+nms0.5+sortdistance-top3*. We observe that the improvements hold: while in all experiments recall reaches 80%, we improve the number of micro-actions necessary to get there from 350 for the basic settings (*init-auto*), to 147 when using NMS during "Add", and to just 86 when also sorting by distance and limiting selection to the top 4 segments.

5.2 Results with human annotators

We now perform several experiments with expert human annotators using the best settings determined in Sec. 5.1: use the machine-generated initialization, and, for the "Add" action, use NMS with

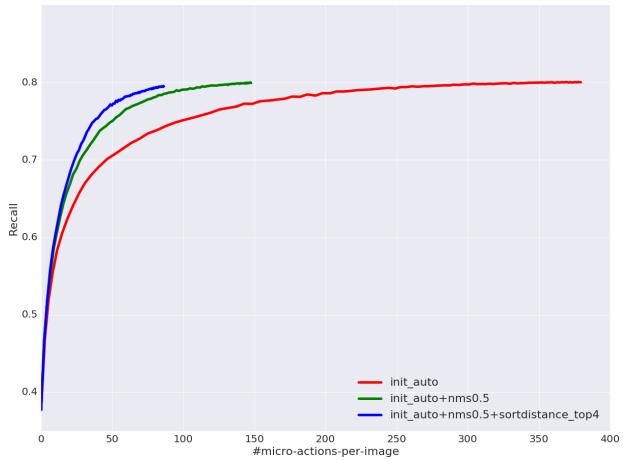


Figure 7: Comparison of our basic settings (red), using NMS within “Add” (green), and using the best settings (blue) on the hold-out set of 4500 images.

Human pixel-wise label agreement		
Fluid annotation vs. COCO+stuff original	Fluid Annotation vs. Polygon annotation	Polygon annotation vs. COCO+stuff original
69%	66%	65%

Table 1: Pixel-wise label agreement across three different annotation methods.

Average annotation time per image		
Fluid annotation	Polygons	COCO+stuff original [11, 51]
175	507	≈ 1140

Table 2: Average annotation time for human experts using the Fluid Annotation interface and the polygon-based interface of LabelMe, and for crowdsourced annotators on COCO+Stuff (16 minutes in [51] plus 3 minutes in [11].)

IoU > 0.5, sort by distance to segment center, and limit selection to the top 4 segments.

Reproducing a reference annotation. We first verify how a human annotator can reproduce a reference annotation, which tests both the flexibility and efficiency of the annotation interface. To do this, we ask two human experts to annotate 25 images while looking at their reference annotation from [11, 51] using two interfaces: (I) Fluid Annotation, and (II) A polygon-based interface representative for what was used to annotate many datasets [17, 57, 88, 92]. More precisely, we use the popular LabelMe [71] as implemented in [1].

We measure quality in terms of recall and efficiency in terms of micro-actions , which allows comparing to our simulation results. For the LabelMe interface [70], each polygon the annotator draws is consider to cost $p + 2$ micro-actions , where p is the number of polygon vertices, plus 2 to input its label (type name and confirm).

Results are shown in Fig. 8. The performance curve for the human annotator behave in a very similar manner to the simulator one. As noted before for the simulator, the recall produced by the human grows rapidly, especially during the first few micro-actions.

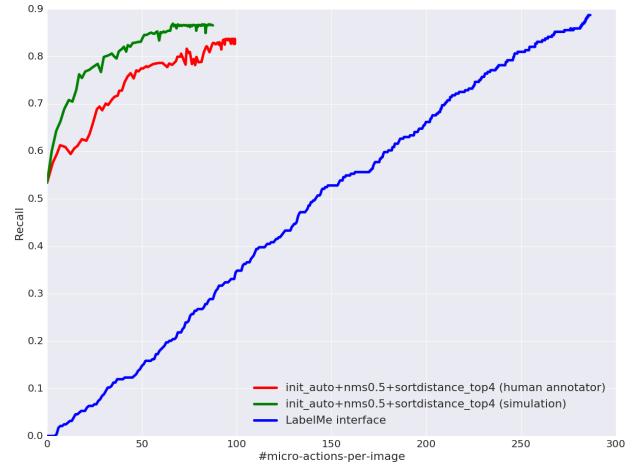


Figure 8: Comparison of annotation results obtained with our Fluid Annotation system (green, red) and with the LabelMe interface (blue) [1, 70].

Moreover, the simulator needs 68 micro-actions per image to produce a recall of 87%, while the human needs 96 micro-actions to get to a slightly lower recall (83%). This similarity in behaviour shows that the simulator is a good proxy for human performance, and that Fluid Annotation is an interface a human can truly efficiently operate.

Importantly, Fluid Annotation is substantially more efficient than the LabelMe interface: using 100 micro-actions , a human annotator produces 35% recall with LabelMe while 83% with Fluid Annotation. From another view, to reach 83% recall, LabelMe requires 2.5× more micro-actions . This demonstrates that Fluid Annotation is a highly effective interface.

Human agreement for different interfaces. While so far we have assumed that the reference annotations [11, 51] are perfect ground-truth, in practice different humans annotating the same image typically produce somewhat different annotations. Therefore, often researchers measure the agreement across multiple annotators [11, 17, 92]. To do this, we annotate 20 images with two expert annotators, this time by showing just the image, without any reference annotation. The first 10 images are annotated by annotator A using Fluid Annotation, and by annotator B using the LabelMe interface. For the second 10 images the annotators switch interfaces: A uses LabelMe and B uses Fluid Annotation. This protocol removes any possible annotator bias from the comparison between interfaces. We measure annotation time and pixel-wise label agreement [11, 17, 92] of all three forms of annotation: Fluid annotation, LabelMe (polygons using our expert annotators), and the process of COCO+stuff [11, 51] (polygons for thing classes and superpixel annotation for stuff classes, all using crowdsourced annotators).

The results are presented in Tab. 1 and 2. All label agreements are relatively close, ranging from 65% to 69%. This level of agreement is reasonable. COCO-stuff [11] reports 74% label agreement on stuff only, where fewer classes can be confused. The authors of the ADE20 dataset [92], which was annotated using LabelMe, report 82% agreement using the same expert annotator six months apart, while 33% agreement between different expert annotators.



Figure 9: Comparison of annotations created using different interfaces: The original COCO+stuff annotations [11, 51], LabelMe [70] polygon annotations, and our Fluid Annotations.

Therefore, we conclude that the quality of annotations for fluid annotation is on par with the compared methods.

Annotation time. Table 2 compares annotation time across annotation interfaces. We see that Fluid Annotation is 3× faster than the LabelMe interface (Table 2). In turn, our annotators with the LabelMe interface are twice as fast as what originally reported for COCO+stuff [11, 51], despite that the bulk of their annotation time was also consumed by drawing polygons. However, this can be attributed to using crowdsourcing versus human experts.

Looking within Fluid Annotation, the action frequencies of an annotator can be broken down as follows: 33% “add”, 31% “remove”, 26% “change label”, and 10% “change depth order”. The “add” action is the most expensive action and takes 18% of the total annotation time. Interestingly, 66% of the total time is spent *between* actions, when the annotator observes the image and decides what to do next.

Qualitative examples. Fig. 9 shows qualitative examples of the various annotation strategies. Generally, Fluid Annotation yields good outlines for most objects. When comparing annotations made by different interfaces, we observe that most of the disagreements are caused by similar segments having a slightly different label.

For example, the wall of the third image is sometimes annotated as *wall-concrete* and sometimes as *wall-other*. As another example, there is legitimate disagreement about how exactly the wooden walk-boards in the second image should be labeled. Finally, some cases are very difficult for Fluid Annotation: in the third image, everything in the doorway on the left is blurry, while the Christmas garland around the door frame is very irregular. As such, the produced segments are only roughly following the actual object boundaries. In fact in these image conditions typically even polygon or superpixel interfaces would produce only approximate outlines.

6 CONCLUSION

We presented Fluid Annotation, an intuitive human-machine collaboration interface for annotating the class label and outline of every object and background region in an image. Fluid annotation substantially reduce human annotation effort, supports full images annotation in a single pass, and it empowers the annotator to choose what to annotate and in which order. We have experimentally demonstrated that Fluid Annotation takes 3× less annotation time than the popular LabelMe interface.

REFERENCES

- [1] [n. d.]. LabelMe: Image Polygonal Annotation with Python. <https://github.com/wkentaro/labelme>
- [2] D. Acuna, H. Ling, A. Kar, and S. Fidler. 2018. Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++. In *CVPR*.
- [3] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. 2014. Multiscale Combinatorial Grouping. In *CVPR*.
- [4] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. 2016. What's the point: Semantic segmentation with point supervision. In *ECCV*.
- [5] S. Bell, P. Upchurch, N. Snavely, and K. Bala. 2015. Material Recognition in the Wild with the Materials in Context Database. In *CVPR*.
- [6] T. Berg and D.A. Forsyth. 2006. Animals on the web. In *CVPR*.
- [7] H. Bilen and A. Vedaldi. 2016. Weakly Supervised Deep Detection Networks. In *CVPR*.
- [8] Arijit Biswas and Devi Parikh. 2013. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*.
- [9] Y. Boykov and M. P. Jolly. 2001. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. In *ICCV*.
- [10] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual recognition with humans in the loop. In *ECCV*.
- [11] H. Caesar, J.R.R. Uijlings, and V. Ferrari. 2018. COCO-Stuff: Thing ans Stuff Classes in Context. In *CVPR*.
- [12] J. Carreira and C. Sminchisescu. 2010. Constrained Parametric Min-Cuts for Automatic Object Segmentation. In *CVPR*.
- [13] L. Castrejón, K. Kundu, R. Urtasun, and S. Fidler. 2017. Annotating object instances with a polygon-rnn. In *CVPR*.
- [14] L.-C. Chen, A. Hermans, F. Schroff, G. Papandreou, P. Wang, and H. Adam. 2017. MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. *ArXiv* (2017).
- [15] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. on PAMI* (2018).
- [16] R.G. Cinbis, J. Verbeek, and C. Schmid. 2014. Multi-fold MIL Training for Weakly Supervised Object Localization. In *CVPR*.
- [17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. *CVPR* (2016).
- [18] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. 2016. Instance-sensitive Fully Convolutional Networks. In *ECCV*.
- [19] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S. Bernstein, Alex Berg, and Li Fei-Fei. 2014. Scalable Multi-label Annotation. In *Proceedings of the 32ND Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 3099–3102.
- [20] T. Deselaers, B. Alexe, and V. Ferrari. 2010. Localizing Objects while Learning Their Appearance. In *ECCV*.
- [21] I. Endres and D. Hoiem. 2014. Category-Independent Object Proposals with Diverse Ranking. *IEEE Trans. on PAMI* 36, 2 (2014), 222–234.
- [22] M. Everingham, S. Eslami, L. van Gool, C. Williams, J. Winn, and A. Zisserman. 2015. The PASCAL Visual Object Classes Challenge: A Retrospective. *IJCV* (2015).
- [23] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. 2010. Object Detection with Discriminatively Trained Part Based Models. *IEEE Trans. on PAMI* 32, 9 (2010).
- [24] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. 2010. Learning Object Categories From Internet Image Searches. In *Proceedings of the IEEE*.
- [25] Y. Freund and R.E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* (1997).
- [26] R. Girshick, J. Donahue, T. Darrell, and J. Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.
- [27] Greg Griffin, Alex Holub, and Pietro Perona. 2007. *The Caltech-256*. Technical Report. Caltech.
- [28] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. 2014. Simultaneous Detection and Segmentation. In *ECCV*.
- [29] M. Häußmann, F.A. Hamprecht, and M. Kandemir. 2017. Variational Bayesian Multiple Instance Learning with Gaussian Processes. In *CVPR*.
- [30] K. He, G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask R-CNN. In *ICCV*.
- [31] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. *CVPR*.
- [32] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*.
- [33] M. Huh, P. Agrawal, and A.A. Efros. 2016. What makes ImageNet good for transfer learning? *NIPS LSCVS workshop*.
- [34] S. Jain and K. Grauman. 2016. Click Carving: Segmenting Objects in Video with Point Clicks. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing*.
- [35] Suyog Dutt Jain and Kristen Grauman. 2013. Predicting sufficient annotation strength for interactive foreground segmentation. In *ICCV*.
- [36] B. Jin, M.V. Ortiz-Segovia, and S. Süsstrunk. 2017. Webly supervised semantic segmentation. In *CVPR*.
- [37] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *CVPR*.
- [38] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. 2010. ContextLocNet: Context-aware Deep Network Models for Weakly Supervised Localization. In *ECCV*.
- [39] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. 2007. Active learning with gaussian processes for object categorization. In *ICCV*.
- [40] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. 2017. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*.
- [41] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. 2018. Panoptic Segmentation. In *ArXiv*.
- [42] A. Kolesnikov and C.H. Lampert. 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*.
- [43] K. Konyushkova, J.R.R. Uijlings, C. Lampert, and V. Ferrari. 2018. Learning Intelligent Dialogs for Bounding Box Annotation. In *CVPR*.
- [44] Adriana Kovashka, Sudheendra Vijayanarasimhan, and Kristen Grauman. 2011. Actively selecting annotations among objects and attributes. In *ICCV*.
- [45] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. 2017. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html* (2017).
- [46] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*.
- [47] A. Li, A. Jabri, A. Joulin, and L. van der Maaten. 2017. Learning Visual N-Grams from Web Data. *ICCV*.
- [48] X. Li, L. Chen, L. Zhang, F. Lin, and W-Y. Ma. 2006. Image Annotation by Large-scale Content-based Image Retrieval. In *ACM Multimedia*.
- [49] J.H. Liew, Y. Wei, W. Xiong, S-H. Ong, and J. Feng. 2017. Regional interactive image segmentation networks. In *ICCV*.
- [50] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. 2016. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *CVPR*.
- [51] T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.
- [52] W. Liu, A. Rabinovich, and A.C. Berg. 2016. ParseNet: Looking Wider to See Better. In *ICLR workshop*.
- [53] J. Long, E. Shelhamer, and T. Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*.
- [54] D. Mahajan, R. Girshick, V. Ramamathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *ArXiv*.
- [55] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. 2018. Deep Extreme Cut: From Extreme Points to Object Segmentation. In *CVPR*.
- [56] Pascal Mettes, Jan C van Gemert, and Cees GM Snoek. 2016. Spot On: Action Localization from Pointily-Supervised Proposals. In *ECCV*.
- [57] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *CVPR*.
- [58] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh. 2013. Analyzing Semantic Segmentation Using Hybrid Human-Machine CRFs. In *CVPR*, 3143–3150.
- [59] N. S. Nagaraja, F. R. Schmidt, and T. Brox. 2015. Video Segmentation with Just a Few Strokes. In *ICCV*.
- [60] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. 2017. Extreme clicking for efficient object annotation. In *ICCV*.
- [61] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. 2017. Training object class detectors with click supervision. In *CVPR*.
- [62] D. P. Papadopoulos, Jasper R. R. Uijlings, F. Keller, and V. Ferrari. 2016. We don't need no bounding-boxes: Training object class detectors using only human verification. In *CVPR*.
- [63] Amar Parkash and Devi Parikh. 2012. Attributes for classifier feedback. In *ECCV*.
- [64] D. Pathak, P. Krähenbühl, and T. Darrell. 2015. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*.
- [65] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, and Hong-Jiang Zhang. 2008. Two-dimensional active learning for image classification. In *CVPR*.
- [66] S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.
- [67] C. Rother, V. Kolmogorov, and A. Blake. 2004. GrabCut: interactive foreground extraction using iterated graph cuts. *SIGGRAPH* 23, 3 (2004), 309–314.
- [68] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* (2015).
- [69] O. Russakovsky, L.-J. Li, and L. Fei-Fei. 2015. Best of both worlds: human-machine collaboration for object annotation. In *CVPR*.

- [70] B. Russel and A. Torralba. 2008. LabelMe: a database and web-based tool for image annotation. *IJCV* 77, 1-3 (2008), 157–173.
- [71] B. C. Russell, K. P. Murphy, and W. T. Freeman. 2008. LabelMe: a database and web-based tool for image annotation. *IJCV* (2008).
- [72] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.
- [73] J. Shotton, J. Winn, C. Rother, and A. Criminisi. 2009. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Appearance, Shape and Context. *IJCV* 81, 1 (2009), 2–23.
- [74] A. Shrivastava, A. Gupta, and R. Girshick. 2016. Training region-based object detectors with online hard example mining. In *CVPR*.
- [75] Behjat Siddique and Abhinav Gupta. 2010. Beyond active noun tagging: Modeling contextual interactions for multi-class active learning. In *CVPR*.
- [76] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.
- [77] H. Su, J. Deng, and L. Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. In *AAAI Human Computation Workshop*.
- [78] C. Sun, A. Shrivastava, S. Singh, and A. Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*.
- [79] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *AAAI*.
- [80] Joseph Tighe and Svetlana Lazebnik. 2013. Superparsing - Scalable Nonparametric Image Parsing with Superpixels. *IJCV* 101, 2 (2013), 329–349.
- [81] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. 2013. Selective search for object recognition. *IJCV* (2013).
- [82] Sudheendra Vijayanarasimhan and Kristen Grauman. 2008. Multi-Level Active Prediction of Useful Image Annotations for Recognition. In *NIPS*.
- [83] Sudheendra Vijayanarasimhan and Kristen Grauman. 2009. What’s it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In *CVPR*.
- [84] Sudheendra Vijayanarasimhan and Kristen Grauman. 2014. Large-scale live active learning: Training object detectors with crawled data and crowds. *IJCV* 108, 1-2 (2014), 97–114.
- [85] Catherine Wah, Grant Van Horn, Steve Branson, Subhrajyoti Maji, Pietro Perona, and Serge Belongie. 2014. Similarity comparisons for interactive fine-grained categorization. In *CVPR*.
- [86] T. Wang, B. Han, and J. Collomosse. 2014. TouchCut: Fast image and video segmentation using single-touch interaction. *CVIU* (2014).
- [87] Z. Wu, C. Shen, and A. van den Hengel. 2016. Bridging Category-level and Instance-level Semantic Image Segmentation. *ArXiv* (2016).
- [88] J. Xiao, K. Ehinger, J. Hays, A. Torralba, and A. Oliva. 2014. SUN Database: Exploring a Large Collection of Scene Categories. *IJCV* (2014), 1–20.
- [89] J. Xu, A. G. Schwing, and R. Urtasun. 2015. Learning to Segment Under Various Forms of Weak Supervision. In *CVPR*.
- [90] N. Xu, B. Price, S. Cohen, J. Yang, and T.S. Huang. 2016. Deep interactive object selection. In *CVPR*.
- [91] Angela Yao, Juergen Gall, Christian Leistner, and Luc Van Gool. 2012. Interactive object detection. In *CVPR*.
- [92] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. 2017. Scene Parsing through ADE20K Dataset. In *CVPR*.
- [93] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao. 2017. Soft Proposal Networks for Weakly Supervised Object Localization. In *ICCV*.