# Exploiting saliency for object segmentation from image level labels

Seong Joon Oh[†]
joon@mpi-inf.mpg.de

Rodrigo Benenson[†]
benenson@mpi-inf.mpg.de

Anna Khoreva[†]
khoreva@mpi-inf.mpg.de

Zeynep Akata[†,‡]
Z.Akata@uva.nl

Mario Fritz[†]
mfritz@mpi-inf.mpg.de

Bernt Schiele[†]
schiele@mpi-inf.mpg.de

[†] Max Planck Institute for Informatics
Saarland Informatics Campus
Saarbrücken, Germany

[‡] Amsterdam Machine Learning Lab
University of Amsterdam
Amsterdam, the Netherlands

## Abstract

*There have been remarkable improvements in the semantic labelling task in the recent years. However, the state of the art methods rely on large-scale pixel-level annotations. This paper studies the problem of training a pixel-wise semantic labeller network from image-level annotations of the present object classes. Recently, it has been shown that high quality seeds indicating discriminative object regions can be obtained from image-level labels. Without additional information, obtaining the full extent of the object is an inherently ill-posed problem due to co-occurrences. We propose using a saliency model as additional information and hereby exploit prior knowledge on the object extent and image statistics. We show how to combine both information sources in order to recover* 80% *of the fully supervised performance – which is the new state of the art in weakly supervised training for pixel-wise semantic labelling.*

## 1. Introduction

Semantic image labelling provides rich information about scenes, but comes at the cost of requiring pixel-wise labelled training data. The accuracy of convnet-based models correlates strongly with the amount of available training data. Collection and annotation of data have become a bottleneck for progress. This problem has raised interest in exploring partially supervised data or different means of supervision, which represents different tradeoffs between annotation efforts and yields in terms of supervision signal for the learning task. For tasks like semantic segmentation there is a need to investigate the minimal supervision to reach the quality comparable to the fully supervised case.


(a)image labels (b)saliency (c)our result
training image test image

Figure 1: We train a semantic labelling network with (a) image-level labels and (b) saliency masks, to generate (c) a pixel-wise labelling of object classes at test time.

A reasonable starting point considers that all training images have image-level labels to indicate the presence or absence of the classes of interest. The weakly supervised learning problem can be seen as a specific instance of learning from constraints [38, 47]. Instead of explicitly supervising the output, the available labels provide a constraint on the desired output. If an image label is absent, no pixel in the image should take that label; if an image label is present at least in one pixel the image must take that label. However, the objects of interest are rarely single pixel. Thus to enforce larger output regions size, shape, or appearance priors are commonly employed (either explicitly or implicitly).

Another reason for exploiting priors, is the fact that the task is fundamentally ambiguous. Strongly co-occurring categories (such as train and rails, sculls and oars, snowbikes and snow) cannot be separated without additional information. Because additional information is needed to solve the task, previous work have explored different avenues, including class-specific size priors [31], crawling additional images [33, 46], or requesting corrections from a human judge [17, 37].

Despite these efforts, the quality of the current best results on the task seems to level out at $\sim 75\%$ of the fully supervised case. Therefore, we argue that additional information sources have to be explored to complement the image level label supervision – in particular addressing the inherent ambiguities of the task. In this work, we propose to exploit class-agnostic saliency as a new ingredient to train for class-specific pixel labelling; and show new state of the art results on Pascal VOC 2012 semantic labelling with image label supervision.

We decompose the problem of object segmentation from image labels into two separate ones: finding the object location (any point on the object), and finding the object's extent. Finding the object extent can be equivalently seen as finding the background area in an image.

For object location we exploit the fact that image classifiers are sensitive to the discriminative areas of an image. Thus, training using the image labels enables to find high confidence points over the objects classes of interest (we call these "object seeds"), as well as high confidence regions for background. A classifier, however, will struggle to delineate the fine details of an object instance, since these might not be particularly discriminative.

For finding the object extent, we exploit the fact that a large portion of photos aim at capturing a subject. Using class-agnostic object saliency we can find the segment corresponding to some of the detected object seeds. Albeit saliency is noisy, it provides information delineating the object extent beyond what seeds can indicate. Our experiments show that this is an effective source of additional information. Our saliency model is itself trained from bounding box annotations only. At no point of our pipeline accurate pixel-wise annotations are used.

In this paper we provide an analysis of the factors that influence the seeds generation, explore the utility of saliency for the task, and report best known results both when using image labels only and image labels with additional data. In summary, our contributions are:

- Propose an effective method for combining seed and saliency for the task of weakly supervised semantic segmentation. Our method achieves the best performance among the known works that utilise image level supervision with or without additional external data.

- Compare recent seed methods side by side, and analyse the importance of saliency towards final quality.

§3 presents our overall architecture, §4 investigates suitable object seeds, and §5 describes how we use saliency to guide the convnet training. Finally §6 discusses the experimental setup, and presents our key results.

## 2. Related work

The last years have seen a renewed interest on weakly supervised training. For semantic labelling, different forms of supervision have been explored: image labels [32, 31, 30, 33, 46, 18], points [3], scribbles [47, 24], and bounding boxes [9, 30, 16]. In this work we focus on image labels as the main form of supervision.

**Object seeds.** Multiple works have considered using a trained classifier (from image level labels) to find areas of the image that belong to a given class, without necessarily enforcing to cover the full object extent (high precision, low recall). Starting from simple strategies such as "probing classifier with different image areas occluded" [50], or back-propagating the class score gradient on the image [41]; significantly more involved strategies have been proposed, mainly by modifying the back-propagation strategy [43, 51, 40], or by solving a per-image optimization problem [6]. All these strategies provide some degree of empirical success but lack a clear theoretical justification, and tend to have rather noisy outputs.
Another approach considers modifying the classifier training procedure so as to have it generate object masks as by-product of a forward-pass. This can be achieved by adding a global a max-pooling [33] or mean-pooling layer [54] in the last stages of the classifier.
In this work we provide an empirical comparison of existing seeders, and explore variants of the mean-pooling approach [54] (§4).

**Pixel labelling from image level supervision.** Initial work approached this problem by adapting multiple-instance learning [32] and expectation-maximization techniques [30], to the semantic labelling case. Without additional priors only poor results are obtained. Using superpixels to inform about the object shape helps [33, 47] and so does using priors on the object size [31]. [18] carefully uses CRFs to propagate the seeds across the image during training, while [36] exploits segment proposals for this.
Most methods compared propose each a new procedure to train a semantic labelling convnet. One exception is [40] which fuses at test time guided back-propagation [43] at multiple convnet layers to generate class-wise heatmaps. They do this over a convnet trained for classification. Being based on classifier, their output masks only partially capture the object extents, as reflected in the comparatively low performance (table 3).
Recognizing the ill-posed nature of the problem, [17] and [37] propose to collect user-feedback as additional information to guide the training of a segmentation convnet.
The closest work to our approach is [46], which also uses saliency as a cue to improve weakly supervised semantic segmentation. There are however a number of differences. First, they use a curriculum learning to expose the segment-

ation convnet first with simple images, and later with more complex ones. We do not need such curriculum, yet reach better results. Second, they use a manually crafted class-agnostic saliency method, while we use a deep learning based one (which provides better cues). Third, their training procedure uses $\sim$ 40k additional images of the classes of interest crawled from the web; we do not use such class-specific external data. Fourth, we report significantly better results, showing in better light the potential of saliency as additional information to guide weakly supervised semantic object labelling.

The seminal work [45] proposed to use "objectness" map from bounding boxes to guide the semantic segmentation. By using bounding boxes, these maps end up being diffuse; in contrast, our saliency map has sharp object boundaries, giving more precise guidance to the semantic labeller.

**Detection boxes from image level supervision.** Detecting object boxes from image labels has similar challenges as pixel labelling. The object location and extent need to be found. State of the art techniques for this task [4, 44, 15] learn to re-score detection proposals using two stream architectures that once trained separate "objectness" scores from class scores. These architecture echo with our approach, where the seeds provide information about the class scores at each pixel (albeit with low recall for foreground classes), and the saliency output provides a per-pixel (class agnostic) "objectness" score.

**Saliency.** Image saliency has multiple connotations, it can refer to a spatial probability map of where a person might look first [48], a probability map of which object a person might look first [23], or a binary mask segmenting the one object a person is most likely to look first [5, 39]. We employ the last definition in this paper. Note that this notion is class-agnostic, and refers more to the composition of the image, than the specific object category.

Like most computer vision areas, hand-crafted methods [14, 28, 8] have now been surpassed by convnet based approaches [53, 22, 21] for object saliency. In this paper we use saliency as an ingredient: improved saliency models would lead to improved results for our method. We describe in §6.1 our saliency model design, trained itself in a weakly supervised fashion from bounding boxes.

**Semantic labelling.** Even when pixel-level annotations are provided (fully supervised case), the task of semantic labelling is far from solved. Multiple convnet architectures have been proposed, including recurrent networks [34], encoder-decoders [29, 1], up-sampling layers [27], using skip layers [2], or dilated convolutions [7, 49], to name a few. Most of them build upon classification architectures such as VGG [42] or ResNet [13]. For comparison with previous work, our experiments are based on the popular DeepLab [7] architecture.

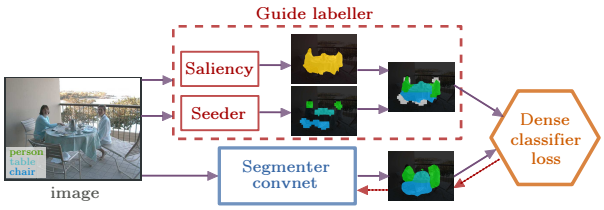



Figure 2: High level Guided Segmentation architecture.

## 3. Guided Segmentation architecture

While previous work have explored sophisticated training losses or involved pipelines, we focus on saliency as an effective prior knowledge, and keep our architecture simple.

We approach the image-level supervised semantic segmentation problem via a system with two modules (see figure 2), we name this architecture "Guided Segmentation". Given an image and image-level labels, the "guide labeller" module combines cues from a seeder (§4) and saliency (§5) sub-modules, producing a rough segmentation mask (the "guide"). Then a segmenter convnet is trained using the produced guide mask as supervision. In this architecture the segmentation convnet is trained in a fully-supervised procedure, using per pixel softmax cross-entropy loss.

In §4 and 5 we explain how we build our guide labeller, by first generating seeds (discriminative areas of objects of interest), and then extending them to better cover the full object extents.

## 4. Finding goods seeds

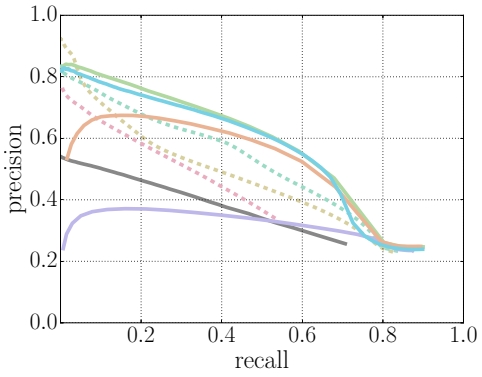
There has been a recent burst of techniques for localising objects from a classifier. Some approaches rely on image gradients from a trained classifier [41, 43, 51], while the others propose to train global average pooling (GAP) based classifiers [54]. Although the classifier based localisation approach has a theoretical limitation that the training objective (image classification) does not match final goal (object locations), they have proved to be effective in practice.

In this section, we review the seeder techniques side by side and compare their empirical performances. We report empirical results on different GAP architectures [54, 18, 7].

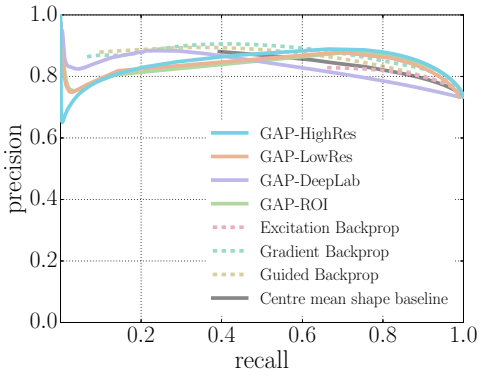
### 4.1. GAP

GAP, or global average pooling layer, can be inserted in the last or penultimate layer of a fully convolutional architecture, which produces a dense prediction, to turn it into a classifier. The resulting architecture is then trained with a classification loss, and at test time the activation maps before the global average pooling layer have been shown to contain localisation information [54].

In our analysis, we consider four different fully convolutional architectures with a GAP layer: `GAP-LowRes`,




(a) Foreground categories




(b) Background category

Figure 3: Precision-recall curves for different seeds. Foreground curves show the average precision and recall among the 20 foreground classes.

`GAP-HighRes`, `GAP-DeepLab`, and `GAP-ROI`. The architectural differences are summarised in table 1, and the full details are provided in the supplementary materials. `GAP-LowRes` [54] is essentially a fully convolutional version of VGG-16 [42]. `GAP-HighRes`, inspired by [18], has 2 times higher output resolution than `GAP-LowRes`. `GAP-DeepLab` is a state of the art semantic segmenter DeepLab with a GAP layer over the dense score output. The main difference between `GAP-HighRes` and `GAP-DeepLab` is the presence of dilated convolutions. `GAP-ROI` is a variant of `GAP-HighRes` where we use the region of interest pooling to replace the sliding window convolutions in the last layers of VGG-16. `GAP-ROI` is identical to `GAP-HighRes`, except for a slight structural variation.

### 4.2. Empirical study

In this section, we empirically compare the seed methods side by side focusing on their utility for the final semantic segmentation task. Together with GAP methods discussed in the previous section, we consider the back-propagation family: Vanilla, Guided, and Excitation back-propagations [41, 43, 51]. We include the centre mean shape baseline that always outputs the average mask shape; it works as a lower bound on the localisation performance.

**Evaluation.** We evaluate each method on the *val* set of the Pascal VOC 2012 [11] segmentation benchmark. We plot the foreground and background precision-recall curves in figure 3. In the foreground case, we compute the mean precision and recall over the 20 Pascal categories.

We define mean precision (mP) as a summary metric for localisation performance. It averages the foreground precision at $20\%$ recall and the background precision at $80\%$ recall; $\mathrm{mP} = \frac{\mathrm{Prec}_{\mathrm{Fg@20\%}} + \mathrm{Prec}_{\mathrm{Bg@80\%}}}{2}$. Intuitively, for the foreground region we only need a small discriminative region, as saliency will fill in the extent; we thus care about precision at $\sim 20\%$ recall. On the other hand, background has more diverse appearance and usually takes a larger region; we thus care about precision at $\sim 80\%$ recall. Since we care about both, we take the average (as for the mAP metric). This metric has shown a good correlation with the final performance in our preliminary experiments.

We measure the classification performance in the standard mean average precision (mAP) metric.

**Implementation details.** We train all four GAP network variants for multi-label image classification over the *trainaug* set of Pascal VOC 2012. Full convnet training details are in the supplementary materials. At test time, we take the output per-class heatmaps before the GAP layer and normalise them by the maximal per-class scores.

For the back-propagation based methods, we obtain image (pseudo-)gradients from the VGG-16 [42] classifier trained on the *trainaug* set of Pascal VOC 2012 (10 582 images in total). We take the maximal absolute gradient value across the RGB channels to generate a rough object mask (following [41]); it is successively smoothed first with vanilla Gaussian kernel and then with dense CRF [19].

In both GAP and backprop variants, we mark pixels with all foreground class scores below $\tau$ as background; other pixels are marked according to the argmax foreground class.

**Results.** See figure 3 for the precision-recall curves. GAP variants have overall greater precision than backprop variants at the same recall rate. We note that the Guided backprop gives highest precision at a very low recall regime ($\sim 5\%$), but the recall is too low to be useful. Among the GAP methods, `GAP-HighRes` and `GAP-ROI` give higher precisions over a wide range of recall. `GAP-DeepLab` shows a significantly lower quality than any other GAP variants.

**Network matters for GAP.** Table 1 shows detailed architectural comparisons and classification/localisation performances of the GAP variants. We observe that the network with higher resolution output has better localisation performance (80.7 mP for `GAP-HighRes` versus 76.5 mP for `GAP-LowRes`). Dilated convolutions significantly hurt the GAP performance (87.0 mP for `GAP-HighRes` versus 57.7

| GAP | -LowRes [54] | -HighRes [18] | -ROI | -DeepLab [7] |
|---|---|---|---|---|
| high res. | ✗ | ✓ | ✓ | ✓ |
| dil. conv. | ✗ | ✗ | ✗ | ✓ |
| ROI pool | ✗ | ✗ | ✓ | ✗ |
| mAP | 88.0 | 87.0 | 87.2 | 92.7 |
| mP | 76.5 | 80.7 | 80.8 | 57.7 |

Table 1: Architectural comparisons among GAP variants together with classification (mAP) and localisation (mP; see text for details) performances. We compare the output resolution (high res.), use of the dilated convolutions (dil. conv.), and the region of interest pooling (ROI pool).

mP for GAP-DeepLab). The architectural choice matters a lot for the localisation performance. This contrasts with the classification performances (mAP), which are stable across design choices. Intriguingly, GAP-DeepLab is in fact the best classifier and the worst seeder at the same time; better design choices for classifiers do not lead to better seeders.

We use GAP-HighRes as the seeder module in the next sections. In [18], foreground and background seeds are handled via different mechanisms; in our experiments we treat all the non-foreground region as background.

## 5. Finding the object extent

Having generated a set of seeds indicating discriminative object areas, the guide labeller needs to find the extent of the object instances (§3).

Without any prior knowledge, it is very hard, if not impossible, to learn the extent of objects only from images and image-level labels. Image-level labels only convey information about commonly occurring patterns that are present in images with positive tags and absent in images with negative tags. The system is thus susceptible to strong inter-class co-occurrences (e.g. train with rail), as well as systematic part occlusions (e.g. feet).

**CRF and CRFLoss.** A traditional approach to make labels match object boundaries is to solve a CRF inference problem [20, 19] over the image grid; where the pair-wise terms relate to the object boundaries. CRF can be applied at three stages: (1) on the seeds (crf-seed), (2) as a loss function during segmenter convnet training (crf-loss) [18], and (3) as a post-processing at test time (crf-postproc). We have experimented with multiple combinations of those (see supplementary materials).

Albeit some gains are observed, these are inconsistent. For example GAP-HighRes and GAP-ROI provide near identical classification and seeding performance (see table 1), yet using the same CRF setup provides +13 mIoU percent points in one, but only +7 pp on the other. In comparison our saliency approach will provide +17 mIoU and +18 mIoU for these two networks respectively (see below).

### 5.1. Saliency

We propose to use object saliency to extract information about the object extent. We work under the assumption that a large portion of the dataset are intentional photographies, which is the case for most datasets crawled from the web such as Pascal [11] and Coco [25]. If the image contains a single label "dog", chances are that the image is about a dog, and that the salient object of the image is a dog. We use a convnet based saliency estimator (detailed in §6.1) which adds the benefit of translation invariance. If two locally salient dogs appear in the image, both will be labelled as foreground.

When using saliency to guide semantic labelling at least two difficulties need to be handled. For one, saliency per-se does not segment object instances. In the example figure 4a, the person-bike is well segmented, but person and bike are not separated. Yet, the ideal Guide labeller (figure 2) should give different labels to these two objects. The second difficulty, clearly visible in the examples of figure 4, is that the salient object might not belong to a category of interest (shirt instead of person in figure 4b) or that the method fails to identify any salient region at all (figure 4c).

We measure the saliency quality when compared to the ground truth foreground on Pascal VOC 2012 validation set. Albeit our convnet saliency model is better than hand-crafted methods [14, 52], in the end only about 20% of images have reasonably good (IoU $> 0.6$) foreground saliency quality. Yet, as we will see in §6, this bit of information is already helpful for the weakly supervised learning task.

Crucially, our saliency system is trained on images containing diverse objects (hundreds of categories), the object categories treated as "unknown". To ensure clean experiments we handicap the system by removing any instance of Pascal categories in the object saliency training set (figure 5). Our saliency model captures a general notion of plausible foreground objects and background areas (details in §6.1).

On every Pascal training image, we obtain a class-agnostic foreground/background binary mask from our saliency model, and high precision/low recall class-specific image labels from the seeds model (§4). We want to combine them in such a way that seed signals are well propagated throughout the foreground saliency mask. We consider two baselines strategies to generate guide labels using saliency but no seeds ($\mathcal{G}_0$ and $\mathcal{G}_1$), and then discuss how we combine saliency with seeds ($\mathcal{G}_2$).

$\mathcal{G}_0$ **Random class assignment.** Given a saliency mask, we assign all foreground pixels to a class randomly picked from the ground truth image labels. If a single "dog" label is present, then all foreground pixels are "dog". Two labels are present ("dog, cat"), then all pixels are either dog or cat.

(a) High quality (b) Medium quality (c) Low quality

Figure 4: Example of our saliency map results on Pascal VOC 2012 data.

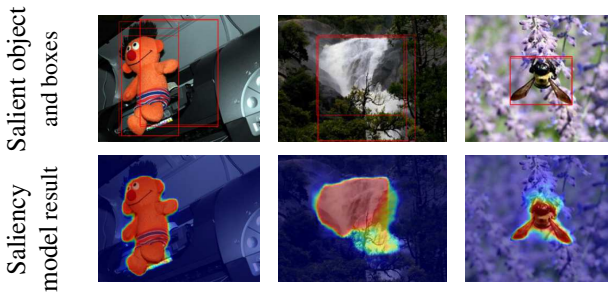



Figure 5: Example of saliency results on its training data. We use MSRA box annotations to train a weakly supervised saliency model. Note that the MSRA subset employed does not contain Pascal categories.

$\mathcal{G}_1$ **Per-connected component classification.** Given a saliency mask, we split it in components, and assign a separate label for each component. The per-component labels are given using a full-image classifier trained using the image labels (classifier details in §6.1). Given a connected component mask $R_i^{fg}$ (with pixel values 1: foreground, 0: background), we compute the classifier scores when feeding the original image ($I$), and when feeding an image with background zeroed ($I \odot R_i^{fg}$). Region $R_i^{fg}$ will be labelled with the ground truth class with the greatest positive score difference before and after zeroing.

$\mathcal{G}_2$ **Propagating seeds.** Here, instead of assigning the label per connected component $R_i^{fg}$ using a classifier, we instead use the seed labels. We also treat the seeds as a set of connected components (seed $R_j^s$). Depending on how the seeds and the foreground regions intersect, we decide the label for each pixel in the guide labeller output.

Our fusion strategy uses five simple ideas. 1) We treat the seeds as reliable small size point predictors of each object instance, but that might leak outside of the object. 2) We assume the saliency might trigger on objects that are not part of the classes of interest. 3) A foreground connected component $R_i^{fg}$ should take the label of the seed touching it, 4) If two (or more) seeds touch the same foreground component, then we want to propagate all the seed labels inside it. 5) When in doubt, mark as ignore. The details for the corner cases are provided in the supplementary material.

Figure 6 provides example results of the different guide strategies. For additional qualitative examples of seeds, saliency foreground, and generated labels, see figure 7. With our guide strategies $\mathcal{G}_0$, $\mathcal{G}_1$, and $\mathcal{G}_2$ at hand, we now proceed to empirically evaluate them in §6.

## 6. Experiments

§6 and 6.1 provide the details of the evaluation and our implementation. §6.2 compares our different guide strategies, and §6.3 compares with previous work on weakly supervised semantic labelling from image-level labels.

**Evaluation.** We evaluate our image-level supervised semantic segmentation system on the Pascal VOC 2012 segmentation benchmark [11]. We report all the intermediate results on the *val* set (1 449 images) and only report the final system result on the *test* set (1 456 images). Evaluation metric is the standard mean intersection-over-union (mIoU).

### 6.1. Implementation details

For training the seeder and segmenter networks, we use the ImageNet [10] pretrained models for initialisation and finetune on the Pascal VOC 2012 *trainaug* set (10 582 images), an extension of the original *train* set (1 464 images) [11, 12]. This is the same procedure used by previous work on fully [7] and weakly supervised learning [18].

**Seeder.** Results in tables 2 and 3 are obtained using `GAP-HighRes` (see §4), trained for image classification on the Pascal trainaug set. The test time foreground threshold $\tau$ is set to 0.2, following the previous literature [54, 18].

$\mathcal{G}_1$ **Classifier.** The guide labeller strategy $\mathcal{G}_1$ uses an image classifier trained on Pascal *trainaug* set. We use the VGG-16 architecture [42] with the softmax cross-entropy multi-label loss.

**Saliency.** Following [53, 22, 21] we re-purpose a semantic labelling network for the task of class-agnostic saliency. We train the DeepLab-v2 ResNet [7] over a subset of MSRA [26], a saliency dataset with *class agnostic* bounding box annotations. We constrain the training only to samples of *non-Pascal* categories. Thus, the saliency model does not leverage class specific features when Pascal images are fed. Out of $25k$ MSRA images, 11 041 remain after filtering.

MRSA provides bounding boxes (from multiple annotators) of the main salient element of each image. To train the saliency model to output pixel-wise masks, we follow [16].

(a) Image (b) Ground truth (c) Seed (d) Saliency (e) $\mathcal{G}_0$ (f) $\mathcal{G}_1$ (g) $\mathcal{G}_2$

Figure 6: Guide labelling strategies example results. The image, its labels ("bicycle, chair"), seeds, and saliency map are their input. White overlay indicates "ignore" pixel label.

We generate segments from the MSRA boxes by applying grabcut over the average box annotation, and use these as supervision for the DeepLab model. The model is trained as a binary semantic labeller for foreground and background regions. The trained model generates masks like the ones shown in figure 5. Although having been trained with images with single salient objects, due to its convolutional nature the network can predict multiple salient regions in the Pascal images (as shown in figure 7).

At test time, the saliency model generates a heatmap of foreground probabilities. We threshold at $50\%$ of the maximal foreground probability to generate the mask.

**Segmenter.** For comparison with previous work we use the DeepLabv1-LargeFOV [7] architecture as our segmenter convnet. The network is trained on Pascal *trainaug* set with $10\,582$ images, using the output of the guide labeller (§2), which uses only the image itself and the presence-absence tags of the 20 Pascal categories as supervision. The network is trained for $8k$ iterations.
Following the standard DeepLab procedure, at test time we up-sample the output to the original image resolution and apply the dense CRF inference [19]. Unless stated otherwise, we use the CRF parameters used for DeepLabv1-LargeFOV [7]. Additional training details and hyper-parameters are given in the supplementary materials.

## 6.2. Ingredients study

Table 2 compares different guide strategies $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$, and oracle versions of $\mathcal{G}_2$. The first row shows the result of training our segmenter using the seeds directly as guide labels. This leads to poor quality (38.7 mIoU). The "Supervision" column shows recall and precision for foreground and background of the guide labels themselves (training data for the segmenter). We can see that the seeds alone have low recall for the foreground ($37\%$). In comparison, using saliency only, $\mathcal{G}_0$ reaches significantly better results, due to the higher foreground recall ($52\%$), at a comparable precision.

Adding a classifier on top of the saliency ($\mathcal{G}_0 \rightarrow \mathcal{G}_1$) provides only a negligible improvement ($45.8 \rightarrow 46.2$). This can be attributed the fact that many Pascal images contain only a single foreground class, and that the classifier might have difficulties recognizing the masked objects. Interestingly, when using a similar classifier to generate seeds instead of scoring the image ($\mathcal{G}_1 \rightarrow \mathcal{G}_2$) we gain 5 pp (percent points, $46.2 \rightarrow 51.2$). This shows that the details of how a classifier is used can make a large difference.

| Method | Seeds | Saliency | Supervision Fg P | Fg R | Bg P | Bg R | val. set mIoU |
|---|---|---|---|---|---|---|---|
| Seeds only | ✓ | ✗ | 69 | 37 | 81 | 95 | 38.7 |
| $\mathcal{G}_0$ | ✗ | ✓ | 65 | 52 | 65 | 52 | 45.8 |
| $\mathcal{G}_1$ | ✗ | ✓ | 75 | 51 | 75 | 51 | 46.2 |
| $\mathcal{G}_2$ | ✓ | ✓ | 73 | 59 | 87 | 95 | 51.2 |
| Saliency oracle | ✓ | ✓ | 89 | 91 | 100 | 99 | 56.9 |

Table 2: Comparison of different guide labeller variants. Pascal VOC 2012 validation set results, without CRF post-processing. Fg/Bg P/R: are foreground/background precision and recall of the guide labels. Discussion in §6.2.

Table 2 also reports a saliency oracle case on top of $\mathcal{G}_2$. If we use the ground truth annotation to generate an ideal saliency mask, we see a significant improvement over $\mathcal{G}_2$ ($51.2 \rightarrow 56.9$). Thus, the quality of saliency is an important ingredient, and there is room for further gains.

## 6.3. Results

Table 3 compares our results with previous related work. We group results by methods that only use ImageNet pre-training and image-level labels (I, P, E; see legend table 3), and methods that use additional data or user-inputs. Here our $\mathcal{G}_0$ and $\mathcal{G}_2$ results include a CRF post-processing (`crf-postproc`). We also experimented with `crf-loss` but did not find a parameter set that provided improved results.

We see that the guide strategies $\mathcal{G}_0$, which uses saliency and random ground-truth label, reaches competitive performance compared to methods using I+P only. This shows that saliency by itself is already a strong cue. Our guide strategy $\mathcal{G}_2$ (which uses seeds and saliency) obtains the best reported results on this task[1]. We even improve over other

[1] [36] also reports 54.3 validation set results, however we do not consider these results comparable since they use the MCG scores [35], which are trained on the ground truth Pascal segments.

| | Method | Data | val. set mIoU | test set mIoU | test set FS% |
|---|---|---|---|---|---|
| Image labels only | MIL-FCN [32] | I+P | 25.0 | 25.6 | 36.5 |
| | CCNN [31] | I+P | 35.3 | 35.6 | 50.6 |
| | WSSL [30] | I+P | 38.2 | 39.6 | 56.3 |
| | MIL+Seg [33] | I+$E_{760k}$ | 42.0 | 40.6 | 57.8 |
| | DCSM [40] | I+P | 44.1 | 45.1 | 64.2 |
| | CheckMask [37] | I+P | 46.6 | - | - |
| | SEC [18] | I+P | 50.7 | 51.7 | 73.5 |
| | AF-ss [36] | I+P | 51.6 | - | - |
| | Seeds only | I+P | 39.8 | - | - |
| More information | CCNN [31] | I+P+Z | - | 45.1 | 64.2 |
| | STC [46] | I+P+S+$E_{40k}$ | 49.8 | 51.2 | 72.8 |
| | CheckMask [37] | I+P+$\mu$ | 51.5 | - | - |
| | MicroAnno [17] | I+P+$\mu$ | 51.9 | 53.2 | 75.7 |
| | $\mathcal{G}_0$ | I+P+S | 48.8 | - | - |
| | $\mathcal{G}_2$ | I+P+S | **55.7** | **56.7** | **80.6** |
| | DeepLabv1 | I+$P_{full}$ | 67.6 | 70.3 | 100 |

Table 3: Comparison of state of the art methods, on Pascal VOC 2012 val. and test set. FS%: fully supervised percent. Ingredients: I: ImageNet classification pre-training, P: Pascal image level tags, $P_{full}$: fully supervised case (pixel wise labels), $E_n$: $n$ extra images with image level tags, S: saliency, Z: per-class size prior, $\mu$: human-in-the-loop micro-annotations.

methods using saliency (STC) or using additional human annotations (MicroAnno, CheckMask). Compared to a fully supervised DeepLabv1 model, our results reach 80% of the fully supervised quality.

## 7. Conclusion

We have addressed the problem of training a semantic segmentation convnet from image labels. Image labels alone can provide high quality seeds, or discriminative object regions, but learning the full object extents is a hard problem. We have shown that saliency is a viable option for feeding the object extent information.

The proposed Guided Segmentation architecture (§3), where the "guide labeller" combines cues from the seeds and saliency, can successfully train a segmentation convnet to achieve the state of the art performance. Our weakly supervised results reach 80% of the fully supervised case.

We expect that a deeper understanding of the seeder methods and improvements on the saliency model can lead to further improvements.


### Acknowledgements

This research was supported by the German Research Foundation (DFG CRC 1223).

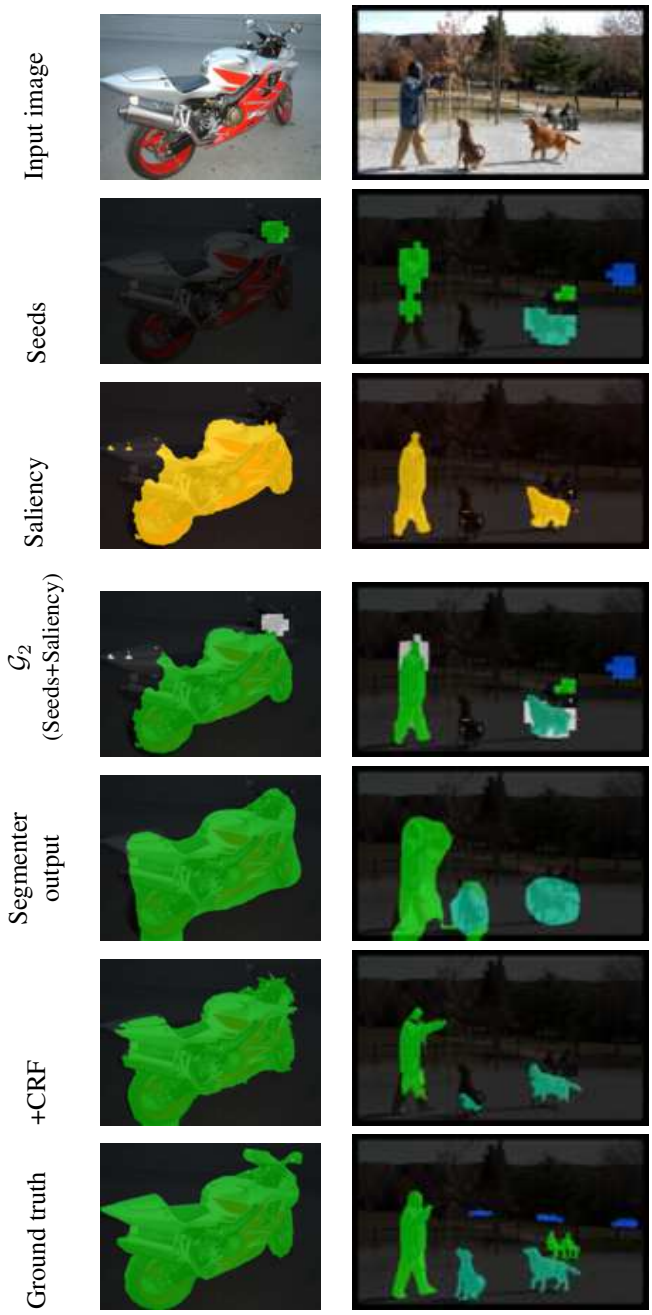




Figure 7: Qualitative examples of the different stages of our system. More examples in the supplementary material.

## References


[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv*, abs/1511.00561, 2015.

[2] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan. Pixelnet: Towards a general pixel-level

architecture. *arXiv preprint arXiv:1609.06694*, 2016.

[3] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. *arXiv preprint arXiv:1506.02106*, 2015.

[4] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[5] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *TIP*, 2015.

[6] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016.

[8] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu. Global contrast based salient region detection. *PAMI*, 2015.

[9] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[12] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

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

[14] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013.

[15] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, 2016.

[16] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele. Weakly supervised semantic labelling and instance segmentation. *arXiv preprint arXiv:1603.07485*, 2016.

[17] A. Kolesnikov and C. Lampert. Improving weakly-supervised object localization by micro-annotation. In *BMVC*, 2016.

[18] A. Kolesnikov and C. H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*. 2011.

[20] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[21] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016.

[22] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *TIP*, 2016.

[23] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[24] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *CVPR*, 2016.

[25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[26] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2011.

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

[28] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *CVPR*, 2013.

[29] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[30] G. Papandreou, L. Chen, K. Murphy, , and A. L. Yuille. Weakly- and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015.

[31] D. Pathak, P. Kraehenbuehl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[32] D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR workshop*, 2015.

[33] P. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional network. In *CVPR*, 2015.

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

[35] J. Pont-Tuset, P. Arbeláez, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *arXiv preprint arXiv:1503.00848*, 2015.

[36] X. Qi, Z. Liu, J. Shi, H. Zhao, and J. Jia. Augmented feedback in semantic segmentation under image level supervision. In *ECCV*, 2016.

[37] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*, 2016.

[38] I. Shcherbatyi and B. Andres. Convexification of learning from constraints. In *GCPR*, 2016.

[39] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *PAMI*, 2016.

[40] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016.

[41] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

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

[43] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR workshop*, 2015.

[44] E. Teh, M. Rochan, and Y. Wang. Attention networks for weakly supervised object localization. In *BMVC*, 2016.

[45] A. Vezhnevets, V. Ferrari, and J. Buhmann. Weakly supervised semantic segmentation with a multi-image model. In *ICCV*, 2011.

[46] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *arXiv preprint arXiv:1509.03150*, 2015.

[47] J. Xu, A. Schwing, and R. Urtasun. Learning to segment under various forms of weak supervision. In *CVPR*, 2015.

[48] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. Can saliency map models predict human egocentric visual attention? In *ACCV*, 2010.

[49] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[50] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[51] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.

[52] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Měch. Minimum barrier salient object detection at 80 fps. In *IEEE International Conference on Computer Vision(ICCV)*, 2015.

[53] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *CVPR*, 2015.

[54] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.