CVPR
VIRTUAL JUNE 19-25

BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation

Weakly supervised segmentation methods using bounding box annotations focus on obtaining a pixel-level mask from each box containing an object. Existing methods typically depend on a class-agnostic mask generator, which operates on the low-level information intrinsic to an image. In this work, we utilize higher-level information from the behavior of a trained object detector, by seeking the smallest areas of the image from which the object detector produces almost the same result as it does from the whole image. These areas constitute a bounding-box attribution map (BBAM), which identifies the target object in its bounding box and thus serves as pseudo ground-truth for weakly supervised semantic and instance segmentation. This approach significantly outperforms recent comparable techniques on both the PASCAL VOC and MS COCO benchmarks in weakly supervised semantic and instance segmentation. In addition, we provide a detailed analysis of our method, offering deeper insight into the behavior of the BBAM.

http://arxiv.org/pdf/2103.08907v1

BoxInst: High-Performance Instance Segmentation With Box Annotations

We present a high-performance method that can achieve mask-level instance segmentation with only bounding-box annotations for training. While this setting has been studied in the literature, here we show significantly stronger performance with a simple design (e.g., dramatically improving previous best reported mask AP of 21.1% in Hsu et al. (2019) to 31.6% on the COCO dataset). Our core idea is to redesign the loss of learning masks in instance segmentation, with no modification to the segmentation network itself. The new loss functions can supervise the mask training without relying on mask annotations. This is made possible with two loss terms, namely, 1) a surrogate term that minimizes the discrepancy between the projections of the ground-truth box and the predicted mask; 2) a pairwise loss that can exploit the prior that proximal pixels with similar colors are very likely to have the same category label. Experiments demonstrate that the redesigned mask loss can yield surprisingly high-quality instance masks with only box annotations. For example, without using any mask annotations, with a ResNet-101 backbone and 3x training schedule, we achieve 33.2% mask AP on COCO test-dev split (vs. 39.1% of the fully supervised counterpart). Our excellent experiment results on COCO and Pascal VOC indicate that our method dramatically narrows the performance gap between weakly and fully supervised instance segmentation. Code is available at: https://git.io/AdelaiDet

http://arxiv.org/pdf/2012.02310v1

Anchor-Constrained Viterbi for Set-Supervised Action Segmentation

This paper is about weakly supervised action segmentation, where the ground truth specifies only a set of actions present in a training video, but not their true temporal ordering. Prior work typically uses a classifier that independently labels video frames for generating the pseudo ground truth, and multiple instance learning for training the classifier. We extend this framework by specifying an HMM, which accounts for co-occurrences of action classes and their temporal lengths, and by explicitly training the HMM on a Viterbi-based loss. Our first contribution is the formulation of a new set-constrained Viterbi algorithm (SCV). Given a video, the SCV generates the MAP action segmentation that satisfies the ground truth. This prediction is used as a framewise pseudo ground truth in our HMM training. Our second contribution in training is a new regularization of feature affinities between training videos that share the same action classes. Evaluation on action segmentation and alignment on the Breakfast, MPII Cooking2, Hollywood Extended datasets demonstrates our significant performance improvement for the two tasks over prior work.

http://arxiv.org/pdf/2002.11925v2

Weakly-Supervised Instance Segmentation via Class-Agnostic Learning With Salient Images

Humans have a strong class-agnostic object segmentation ability and can outline boundaries of unknown objects precisely, which motivates us to propose a box-supervised class-agnostic object segmentation (BoxCaseg) based solution for weakly-supervised instance segmentation. The BoxCaseg model is jointly trained using box-supervised images and salient images in a multi-task learning manner. The fine-annotated salient images provide class-agnostic and precise object localization guidance for box-supervised images. The object masks predicted by a pretrained BoxCaseg model are refined via a novel merged and dropped strategy as proxy ground truth to train a Mask R-CNN for weakly-supervised instance segmentation. Only using 7991 salient images, the weakly-supervised Mask R-CNN is on par with fully-supervised Mask R-CNN on PASCAL VOC and significantly outperforms previous state-of-the-art box-supervised instance segmentation methods on COCO. The source code, pretrained models and datasets are available at \url{https://github.com/hustvl/BoxCaseg}.

http://arxiv.org/pdf/2104.01526v1

3D Spatial Recognition Without Spatially Labeled 3D

We introduce WyPR, a Weakly-supervised framework for Point cloud Recognition, requiring only scene-level class tags as supervision. WyPR jointly addresses three core 3D recognition tasks: point-level semantic segmentation, 3D proposal generation, and 3D object detection, coupling their predictions through self and cross-task consistency losses. We show that in conjunction with standard multiple-instance learning objectives, WyPR can detect and segment objects in point cloud data without access to any spatial labels at training time. We demonstrate its efficacy using the ScanNet and S3DIS datasets, outperforming prior state of the art on weakly-supervised segmentation by more than 6% mIoU. In addition, we set up the first benchmark for weakly-supervised 3D object detection on both datasets, where WyPR outperforms standard approaches and establishes strong baselines for future work.

http://arxiv.org/pdf/2105.06461v1

Weakly Supervised Instance Segmentation for Videos With Temporal Mask Consistency

Weakly supervised instance segmentation reduces the cost of annotations required to train models. However, existing approaches which rely only on image-level class labels predominantly suffer from errors due to (a) partial segmentation of objects and (b) missing object predictions. We show that these issues can be better addressed by training with weakly labeled videos instead of images. In videos, motion and temporal consistency of predictions across frames provide complementary signals which can help segmentation. We are the first to explore the use of these video signals to tackle weakly supervised instance segmentation. We propose two ways to leverage this information in our model. First, we adapt inter-pixel relation network (IRN) to effectively incorporate motion information during training. Second, we introduce a new MaskConsist module, which addresses the problem of missing object instances by transferring stable predictions between neighboring frames during training. We demonstrate that both approaches together improve the instance segmentation metric AP_{50} on video frames of two datasets: Youtube-VIS and Cityscapes by 5% and 3% respectively.

http://arxiv.org/pdf/2103.12886v1

Weakly Supervised Instance Segmentation by Deep Community Learning

We present a weakly supervised instance segmentation algorithm based on deep community learning with multiple tasks. This task is formulated as a combination of weakly supervised object detection and semantic segmentation, where individual objects of the same class are identified and segmented separately. We address this problem by designing a unified deep neural network architecture, which has a positive feedback loop of object detection with bounding box regression, instance mask generation, instance segmentation, and feature extraction. Each component of the network makes active interactions with others to improve accuracy, and the end-to-end trainability of our model makes our results more robust and reproducible. The proposed algorithm achieves state-of-the-art performance in the weakly supervised setting without any additional training such as Fast R-CNN and Mask R-CNN on the standard benchmark dataset. The implementation of our algorithm is available on the project webpage: https://cv.snu.ac.kr/research/WSIS_CL.

```
InProceedings{Hwang_2021_WACV,
    author    = {Hwang, Jaedong and Kim, Seohyun and Son, Jeany and Han, Bohyung},
    title     = {Weakly Supervised Instance Segmentation by Deep Community Learning},
    booktitle = {Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)},
    month     = {January},
    year      = {2021},
    pages     = {1020-1029}
}
```

https://openaccess.thecvf.com/content/WACV2021/papers/Hwang_Weakly_Supervised_Instance_Segmentation_by_Deep_Community_Learning_WACV_2021_paper.pdf

AdaCoSeg: Adaptive Shape Co-Segmentation With Group Consistency Loss

We introduce AdaCoSeg, a deep neural network architecture for adaptive co-segmentation of a set of 3D shapes represented as point clouds. Differently from the familiar single-instance segmentation problem, co-segmentation is intrinsically contextual: how a shape is segmented can vary depending on the set it is in. Hence, our network features an adaptive learning module to produce a consistent shape segmentation which adapts to a set. Specifically, given an input set of unsegmented shapes, we first employ an offline pre-trained part prior network to propose per-shape parts. Then the co-segmentation network iteratively and jointly optimizes the part labelings across the set subjected to a novel group consistency loss defined by matrix ranks. While the part prior network can be trained with noisy and inconsistently segmented shapes, the final output of AdaSeg is a consistent part labeling for the input set, with each shape segmented into up to (a user-specified) K parts. Overall, our method is weakly supervised, producing segmentations tailored to the test set, without consistent ground-truth segmentations. We show qualitative and quantitative results from AdaSeg and evaluate it via ablation studies and comparisons to state-of-the-art co-segmentation methods.

```
InProceedings{Zhu_2020_CVPR,
author = {Zhu, Chenyang and Xu, Kai and Chaudhuri, Siddhartha and Yi, Li and Guibas, Leonidas J. and Zhang, Hao},
title = {AdaCoSeg: Adaptive Shape Co-Segmentation With Group Consistency Loss},
booktitle = {Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2020}
}
```

https://openaccess.thecvf.com/content_CVPR_2020/papers/Zhu_AdaCoSeg_Adaptive_Shape_Co-Segmentation_With_Group_Consistency_Loss_CVPR_2020_paper.pdf

Learning Video Object Segmentation From Unlabeled Videos

We propose a new method for video object segmentation (VOS) that addresses object pattern learning from unlabeled videos, unlike most existing methods which rely heavily on extensive annotated data. We introduce a unified unsupervised/weakly supervised learning framework, called MuG, that comprehensively captures intrinsic properties of VOS at multiple granularities. Our approach can help advance understanding of visual patterns in VOS and significantly reduce annotation burden. With a carefully-designed architecture and strong representation learning ability, our learned model can be applied to diverse VOS settings, including object-level zero-shot VOS, instance-level zero-shot VOS, and one-shot VOS. Experiments demonstrate promising performance in these settings, as well as the potential of MuG in leveraging unlabeled data to further improve the segmentation accuracy.

```
InProceedings{Lu_2020_CVPR,
author = {Lu, Xiankai and Wang, Wenguan and Shen, Jianbing and Tai, Yu-Wing and Crandall, David J. and Hoi, Steven C. H.},
title = {Learning Video Object Segmentation From Unlabeled Videos},
booktitle = {Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2020}
}
```

https://openaccess.thecvf.com/content_CVPR_2020/papers/Lu_Learning_Video_Object_Segmentation_From_Unlabeled_Videos_CVPR_2020_paper.pdf

Set-Constrained Viterbi for Set-Supervised Action Segmentation

This paper is about weakly supervised action segmentation, where the ground truth specifies only a set of actions present in a training video, but not their true temporal ordering. Prior work typically uses a classifier that independently labels video frames for generating the pseudo ground truth, and multiple instance learning for training the classifier. We extend this framework by specifying an HMM, which accounts for co-occurrences of action classes and their temporal lengths, and by explicitly training the HMM on a Viterbi-based loss. Our first contribution is the formulation of a new set-constrained Viterbi algorithm (SCV). Given a video, the SCV generates the MAP action segmentation that satisfies the ground truth. This prediction is used as a framewise pseudo ground truth in our HMM training. Our second contribution in training is a new regularization of feature affinities between training videos that share the same action classes. Evaluation on action segmentation and alignment on the Breakfast, MPII Cooking2, Hollywood Extended datasets demonstrates our significant performance improvement for the two tasks over prior work.

```
InProceedings{Li_2020_CVPR,
author = {Li, Jun and Todorovic, Sinisa},
title = {Set-Constrained Viterbi for Set-Supervised Action Segmentation},
booktitle = {Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2020}
}
```

https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Set-Constrained_Viterbi_for_Set-Supervised_Action_Segmentation_CVPR_2020_paper.pdf

Causal Intervention for Weakly-Supervised Semantic Segmentation

We present a causal inference framework to improve Weakly-Supervised Semantic Segmentation (WSSS). Specifically, we aim to generate better pixel-level pseudo-masks by using only image-level labels -- the most crucial step in WSSS. We attribute the cause of the ambiguous boundaries of pseudo-masks to the confounding context, e.g., the correct image-level classification of "horse" and "person" may be not only due to the recognition of each instance, but also their co-occurrence context, making the model inspection (e.g., CAM) hard to distinguish between the boundaries. Inspired by this, we propose a structural causal model to analyze the causalities among images, contexts, and class labels. Based on it, we develop a new method: Context Adjustment (CONTA), to remove the confounding bias in image-level classification and thus provide better pseudo-masks as ground-truth for the subsequent segmentation model. On PASCAL VOC 2012 and MS-COCO, we show that CONTA boosts various popular WSSS methods to new state-of-the-arts.

```
inproceedings{NEURIPS2020_07211688,
 author = {Zhang, Dong and Zhang, Hanwang and Tang, Jinhui and Hua, Xian-Sheng and Sun, Qianru},
 booktitle = {Advances in Neural Information Processing Systems},
 editor = {H. Larochelle and M. Ranzato and R. Hadsell and M. F. Balcan and H. Lin},
 pages = {655--666},
 publisher = {Curran Associates, Inc.},
 title = {Causal Intervention for Weakly-Supervised Semantic Segmentation},
 url = {https://proceedings.neurips.cc/paper/2020/file/07211688a0869d995947a8fb11b215d6-Paper.pdf},
 volume = {33},
 year = {2020}
}
```

https://papers.nips.cc/paper/2020/file/07211688a0869d995947a8fb11b215d6-Paper.pdf

Weakly Supervised Instance Segmentation by Learning Annotation Consistent Instances

Recent approaches for weakly supervised instance segmentations depend on two components: (i) a pseudo label generation model that provides instances which are consistent with a given annotation; and (ii) an instance segmentation model, which is trained in a supervised manner using the pseudo labels as ground-truth. Unlike previous approaches, we explicitly model the uncertainty in the pseudo label generation process using a conditional distribution. The samples drawn from our conditional distribution provide accurate pseudo labels due to the use of semantic class aware unary terms, boundary aware pairwise smoothness terms, and annotation aware higher order terms. Furthermore, we represent the instance segmentation model as an annotation agnostic prediction distribution. In contrast to previous methods, our representation allows us to define a joint probabilistic learning objective that minimizes the dissimilarity between the two distributions. Our approach achieves state of the art results on the PASCAL VOC 2012 data set, outperforming the best baseline by 4.2% mAP@0.5 and 4.8% mAP@0.75.

https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123730256.pdf

Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations

This paper presents a novel approach for learning instance segmentation with image-level class labels as supervision. Our approach generates pseudo instance segmentation labels of training images, which are used to train a fully supervised model. For generating the pseudo labels, we first identify confident seed areas of object classes from attention maps of an image classification model, and propagate them to discover the entire instance areas with accurate boundaries. To this end, we propose IRNet, which estimates rough areas of individual instances and detects boundaries between different object classes. It thus enables to assign instance labels to the seeds and to propagate them within the boundaries so that the entire areas of instances can be estimated accurately. Furthermore, IRNet is trained with inter-pixel relations on the attention maps, thus no extra supervision is required. Our method with IRNet achieves an outstanding performance on the PASCAL VOC 2012 dataset, surpassing not only previous state-of-the-art trained with the same level of supervision, but also some of previous models relying on stronger supervision.

```
InProceedings{Ahn_2019_CVPR,
author = {Ahn, Jiwoon and Cho, Sunghyun and Kwak, Suha},
title = {Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations},
booktitle = {Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2019}
}
```

https://openaccess.thecvf.com/content_CVPR_2019/papers/Ahn_Weakly_Supervised_Learning_of_Instance_Segmentation_With_Inter-Pixel_Relations_CVPR_2019_paper.pdf

Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification From the Bottom Up

Given a training dataset composed of images and corresponding category labels, deep convolutional neural networks show a strong ability in mining discriminative parts for image classification. However, deep convolutional neural networks trained with image level labels only tend to focus on the most discriminative parts while missing other object parts, which could provide complementary information. In this paper, we approach this problem from a different perspective. We build complementary parts models in a weakly supervised manner to retrieve information suppressed by dominant object parts detected by convolutional neural networks. Given image level labels only, we first extract rough object instances by performing weakly supervised object detection and instance segmentation using Mask R-CNN and CRF-based segmentation. Then we estimate and search for the best parts model for each object instance under the principle of preserving as much diversity as possible. In the last stage, we build a bi-directional long short-term memory (LSTM) network to fuze and encode the partial information of these complementary parts into a comprehensive feature for image classification. Experimental results indicate that the proposed method not only achieves significant improvement over our baseline models, but also outperforms state-of-the-art algorithms by a large margin (6.7%, 2.8%, 5.2% respectively) on Stanford Dogs 120, Caltech-UCSD Birds 2011-200 and Caltech 256.

```
InProceedings{Ge_2019_CVPR,
author = {Ge, Weifeng and Lin, Xiangru and Yu, Yizhou},
title = {Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification From the Bottom Up},
booktitle = {Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2019}
}
```

https://openaccess.thecvf.com/content_CVPR_2019/papers/Ge_Weakly_Supervised_Complementary_Parts_Models_for_Fine-Grained_Image_Classification_From_CVPR_2019_paper.pdf

Learning Instance Activation Maps for Weakly Supervised Instance Segmentation

Discriminative region responses residing inside an object instance can be extracted from networks trained with image-level label supervision. However, learning the full extent of pixel-level instance response in a weakly supervised manner remains unexplored. In this work, we tackle this challenging problem by using a novel instance extent filling approach. We first design a process to selectively collect pseudo supervision from noisy segment proposals obtained with previously published techniques. The pseudo supervision is used to learn a differentiable filling module that predicts a class-agnostic activation map for each instance given the image and an incomplete region response. We refer to the above maps as Instance Activation Maps (IAMs), which provide a fine-grained instance-level representation and allow instance masks to be extracted by lightweight CRF. Extensive experiments on the PASCAL VOC12 dataset show that our approach beats the state-of-the-art weakly supervised instance segmentation methods by a significant margin and increases the inference speed by an order of magnitude. Our method also generalizes well across domains and to unseen object categories. Without fine-tuning for the specific tasks, our model trained on VOC12 dataset (20 classes) obtains top performance for weakly supervised object localization on the CUB dataset (200 classes) and achieves competitive results on three widely used salient object detection benchmarks.

```
InProceedings{Zhu_2019_CVPR,
author = {Zhu, Yi and Zhou, Yanzhao and Xu, Huijuan and Ye, Qixiang and Doermann, David and Jiao, Jianbin},
title = {Learning Instance Activation Maps for Weakly Supervised Instance Segmentation},
booktitle = {Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2019}
}
```

https://openaccess.thecvf.com/content_CVPR_2019/papers/Zhu_Learning_Instance_Activation_Maps_for_Weakly_Supervised_Instance_Segmentation_CVPR_2019_paper.pdf

Learning Segmentation Masks with the Independence Prior

An instance with a bad mask might make a composite image that uses it look fake. This encourages us to learn segmentation by generating realistic composite images. To achieve this, we propose a novel framework that exploits a new proposed prior called the independence prior based on Generative Adversarial Networks (GANs). The generator produces an image with multiple category-specific instance providers, a layout module and a composition module. Firstly, each provider independently outputs a category-specific instance image with a soft mask. Then the provided instances' poses are corrected by the layout module. Lastly, the composition module combines these instances into a final image. Training with adversarial loss and penalty for mask area, each provider learns a mask that is as small as possible but enough to cover a complete category-specific instance. Weakly supervised semantic segmentation methods widely use grouping cues modeling the association between image parts, which are either artificially designed or learned with costly segmentation labels or only modeled on local pairs. Unlike them, our method automatically models the dependence between any parts and learns instance segmentation. We apply our framework in two cases: (1) Foreground segmentation on category-specific images with box-level annotation. (2) Unsupervised learning of instance appearances and masks with only one image of homogeneous object cluster (HOC). We get appealing results in both tasks, which shows the independence prior is useful for instance segmentation and it is possible to unsupervisedly learn instance masks with only one image.

http://arxiv.org/pdf/1811.04682v2

Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior

This paper presents a weakly supervised instance segmentation method that consumes training data with tight bounding box annotations. The major difficulty lies in the uncertain figure-ground separation within each bounding box since there is no supervisory signal about it. We address the difficulty by formulating the problem as a multiple instance learning (MIL) task, and generate positive and negative bags based on the sweeping lines of each bounding box. The proposed deep model integrates MIL into a fully supervised instance segmentation network, and can be derived by the objective consisting of two terms, i.e., the unary term and the pairwise term. The former estimates the foreground and background areas of each bounding box while the latter maintains the unity of the estimated object masks. The experimental results show that our method performs favorably against existing weakly supervised methods and even surpasses some fully supervised methods for instance segmentation on the PASCAL VOC dataset.

```
@inproceedings{NEURIPS2019_e6e71329,
 author = {Hsu, Cheng-Chun and Hsu, Kuang-Jui and Tsai, Chung-Chi and Lin, Yen-Yu and Chuang, Yung-Yu},
 booktitle = {Advances in Neural Information Processing Systems},
 editor = {H. Wallach and H. Larochelle and A. Beygelzimer and F. d\textquotesingle Alch\'{e}-Buc and E. Fox and R. Garnett},
 pages = {},
 publisher = {Curran Associates, Inc.},
 title = {Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior},
 url = {https://proceedings.neurips.cc/paper/2019/file/e6e713296627dff6475085cc6a224464-Paper.pdf},
 volume = {32},
 year = {2019}
}
```

https://papers.nips.cc/paper/2019/file/e6e713296627dff6475085cc6a224464-Paper.pdf

InstaBoost: Boosting Instance Segmentation via Probability Map Guided Copy-Pasting

Instance segmentation requires a large number of training samples to achieve satisfactory performance and benefits from proper data augmentation. To enlarge the training set and increase the diversity, previous methods have investigated using data annotation from other domain (e.g. bbox, point) in a weakly supervised mechanism. In this paper, we present a simple, efficient and effective method to augment the training set using the existing instance mask annotations. Exploiting the pixel redundancy of the background, we are able to improve the performance of Mask R-CNN for 1.7 mAP on COCO dataset and 3.3 mAP on Pascal VOC dataset by simply introducing random jittering to objects. Furthermore, we propose a location probability map based approach to explore the feasible locations that objects can be placed based on local appearance similarity. With the guidance of such map, we boost the performance of R101-Mask R-CNN on instance segmentation from 35.7 mAP to 37.9 mAP without modifying the backbone or network structure. Our method is simple to implement and does not increase the computational complexity. It can be integrated into the training pipeline of any instance segmentation model without affecting the training and inference efficiency. Our code and models have been released at https://github.com/GothicAi/InstaBoost.

```
InProceedings{Fang_2019_ICCV,
author = {Fang, Hao-Shu and Sun, Jianhua and Wang, Runzhong and Gou, Minghao and Li, Yong-Lu and Lu, Cewu},
title = {InstaBoost: Boosting Instance Segmentation via Probability Map Guided Copy-Pasting},
booktitle = {Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)},
month = {October},
year = {2019}
}
```

https://openaccess.thecvf.com/content_ICCV_2019/papers/Fang_InstaBoost_Boosting_Instance_Segmentation_via_Probability_Map_Guided_Copy-Pasting_ICCV_2019_paper.pdf

ICCV 2019
Seoul, Korea

Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation

Weakly-supervised instance segmentation aims to detect and segment object instances precisely, given image-level labels only. Unlike previous methods which are composed of multiple offline stages, we propose Sequential Label Propagation and Enhancement Networks (referred as Label-PEnet) that progressively transforms image-level labels to pixel-wise labels in a coarse-to-fine manner. We design four cascaded modules including multi-label classification, object detection, instance refinement and instance segmentation, which are implemented sequentially by sharing the same backbone. The cascaded pipeline is trained alternatively with a curriculum learning strategy that generalizes labels from high level images to low-level pixels gradually with increasing accuracy. In addition, we design a proposal calibration module to explore the ability of classification networks to find key pixels that identify object parts, which serves as a post validation strategy running in the inverse order. We evaluate the efficiency of our Label-PEnet in mining instance masks on standard benchmarks: PASCAL VOC 2007 and 2012. Experimental results show that Label-PEnet outperforms the state-of-art algorithms by a clear margin, and obtains comparable performance even with fully supervised approaches.

```
InProceedings{Ge_2019_ICCV,
author = {Ge, Weifeng and Guo, Sheng and Huang, Weilin and Scott, Matthew R.},
title = {Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation},
booktitle = {Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)},
month = {October},
year = {2019}
}
```

https://openaccess.thecvf.com/content_ICCV_2019/papers/Ge_Label-PEnet_Sequential_Label_Propagation_and_Enhancement_Networks_for_Weakly_Supervised_ICCV_2019_paper.pdf

Weakly Supervised Object Detection With Segmentation Collaboration

Weakly supervised object detection aims at learning precise object detectors, given image category labels. In recent prevailing works, this problem is generally formulated as a multiple instance learning module guided by an image classification loss. The object bounding box is assumed to be the one contributing most to the classification among all proposals. However, the region contributing most is also likely to be a crucial part or the supporting context of an object. To obtain a more accurate detector, in this work we propose a novel end-to-end weakly supervised detection approach, where a newly introduced generative adversarial segmentation module interacts with the conventional detection module in a collaborative loop. The collaboration mechanism takes full advantages of the complementary interpretations of the weakly supervised localization task, namely detection and segmentation tasks, forming a more comprehensive solution. Consequently, our method obtains more precise object bounding boxes, rather than parts or irrelevant surroundings. Expectedly, the proposed method achieves an accuracy of 53.7% on the PASCAL VOC 2007 dataset, outperforming the state-of-the-arts and demonstrating its superiority for weakly supervised object detection.

```
InProceedings{Li_2019_ICCV,
author = {Li, Xiaoyan and Kan, Meina and Shan, Shiguang and Chen, Xilin},
title = {Weakly Supervised Object Detection With Segmentation Collaboration},
booktitle = {Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)},
month = {October},
year = {2019}
}
```

https://openaccess.thecvf.com/content_ICCV_2019/papers/Li_Weakly_Supervised_Object_Detection_With_Segmentation_Collaboration_ICCV_2019_paper.pdf

C-MIDN: Coupled Multiple Instance Detection Network With Segmentation Guidance for Weakly Supervised Object Detection

Weakly supervised object detection (WSOD) that only needs image-level annotations has obtained much attention recently. By combining convolutional neural network with multiple instance learning method, Multiple Instance Detection Network (MIDN) has become the most popular method to address the WSOD problem and been adopted as the initial model in many works. We argue that MIDN inclines to converge to the most discriminative object parts, which limits the performance of methods based on it. In this paper, we propose a novel Coupled Multiple Instance Detection Network (C-MIDN) to address this problem. Specifically, we use a pair of MIDNs, which work in a complementary manner with proposal removal. The localization information of the MIDNs is further coupled to obtain tighter bounding boxes and localize multiple objects. We also introduce a Segmentation Guided Proposal Removal (SGPR) algorithm to guarantee the MIL constraint after the removal and ensure the robustness of C-MIDN. Through a simple implementation of the C-MIDN with online detector refinement, we obtain 53.6% and 50.3% mAP on the challenging PASCAL VOC 2007 and 2012 benchmarks respectively, which significantly outperform the previous state-of-the-arts.

```
InProceedings{Gao_2019_ICCV,
author = {Gao, Yan and Liu, Boxiao and Guo, Nan and Ye, Xiaochun and Wan, Fang and You, Haihang and Fan, Dongrui},
title = {C-MIDN: Coupled Multiple Instance Detection Network With Segmentation Guidance for Weakly Supervised Object Detection},
booktitle = {Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)},
month = {October},
year = {2019}
}
```

https://openaccess.thecvf.com/content_ICCV_2019/papers/Gao_C-MIDN_Coupled_Multiple_Instance_Detection_Network_With_Segmentation_Guidance_for_ICCV_2019_paper.pdf

CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentation

Histopathology image analysis plays a critical role in cancer diagnosis and treatment. To automatically segment the cancerous regions, fully supervised segmentation algorithms require labor-intensive and time-consuming labeling at the pixel level. In this research, we propose CAMEL, a weakly supervised learning framework for histopathology image segmentation using only image-level labels. Using multiple instance learning (MIL)-based label enrichment, CAMEL splits the image into latticed instances and automatically generates instance-level labels. After label enrichment, the instance-level labels are further assigned to the corresponding pixels, producing the approximate pixel-level labels and making fully supervised training of segmentation models possible. CAMEL achieves comparable performance with the fully supervised approaches in both instance-level classification and pixel-level segmentation on CAMELYON16 and a colorectal adenoma dataset. Moreover, the generality of the automatic labeling methodology may benefit future weakly supervised learning studies for histopathology image analysis.

```
InProceedings{Xu_2019_ICCV,
author = {Xu, Gang and Song, Zhigang and Sun, Zhuo and Ku, Calvin and Yang, Zhe and Liu, Cancheng and Wang, Shuhao and Ma, Jianpeng and Xu, Wei},
title = {CAMEL: A Weakly Supervised Learning Framework for Histopathology Image Segmentation},
booktitle = {Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)},
month = {October},
year = {2019}
}
```

https://openaccess.thecvf.com/content_ICCV_2019/papers/Xu_CAMEL_A_Weakly_Supervised_Learning_Framework_for_Histopathology_Image_Segmentation_ICCV_2019_paper.pdf

Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning

Supervised object detection and semantic segmentation require object or even pixel level annotations. When there exist image level labels only, it is challenging for weakly supervised algorithms to achieve accurate predictions. The accuracy achieved by top weakly supervised algorithms is still significantly lower than their fully supervised counterparts. In this paper, we propose a novel weakly supervised curriculum learning pipeline for multi-label object recognition, detection and semantic segmentation. In this pipeline, we first obtain intermediate object localization and pixel labeling results for the training images, and then use such results to train task-specific deep networks in a fully supervised manner. The entire process consists of four stages, including object localization in the training images, filtering and fusing object instances, pixel labeling for the training images, and task-specific network training. To obtain clean object instances in the training images, we propose a novel algorithm for filtering, fusing and classifying object instances collected from multiple solution mechanisms. In this algorithm, we incorporate both metric learning and density-based clustering to filter detected object instances. Experiments show that our weakly supervised pipeline achieves state-of-the-art results in multi-label image classification as well as weakly supervised object detection and very competitive results in weakly supervised semantic segmentation on MS-COCO, PASCAL VOC 2007 and PASCAL VOC 2012.

```
InProceedings{Ge_2018_CVPR,
author = {Ge, Weifeng and Yang, Sibei and Yu, Yizhou},
title = {Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning},
booktitle = {Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2018}
}
```

https://openaccess.thecvf.com/content_cvpr_2018/papers/Ge_Multi-Evidence_Filtering_and_CVPR_2018_paper.pdf

Weakly Supervised Instance Segmentation Using Class Peak Response

Weakly supervised instance segmentation with image-level labels, instead of expensive pixel-level masks, remains unexplored. In this paper, we tackle this challenging problem by exploiting class peak responses to enable a classification network for instance mask extraction. With image labels supervision only, CNN classifiers in a fully convolutional manner can produce class response maps, which specify classification confidence at each image location. We observed that local maximums, i.e., peaks, in a class response map typically correspond to strong visual cues residing inside each instance. Motivated by this, we first design a process to stimulate peaks to emerge from a class response map. The emerged peaks are then back-propagated and effectively mapped to highly informative regions of each object instance, such as instance boundaries. We refer to the above maps generated from class peak responses as Peak Response Maps (PRMs). PRMs provide a fine-detailed instance-level representation, which allows instance masks to be extracted even with some off-the-shelf methods. To the best of our knowledge, we for the first time report results for the challenging image-level supervised instance segmentation task. Extensive experiments show that our method also boosts weakly supervised pointwise localization as well as semantic segmentation performance, and reports state-of-the-art results on popular benchmarks, including PASCAL VOC 2012 and MS COCO.

```
InProceedings{Zhou_2018_CVPR,
author = {Zhou, Yanzhao and Zhu, Yi and Ye, Qixiang and Qiu, Qiang and Jiao, Jianbin},
title = {Weakly Supervised Instance Segmentation Using Class Peak Response},
booktitle = {Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2018}
}
```

https://openaccess.thecvf.com/content_cvpr_2018/papers/Zhou_Weakly_Supervised_Instance_CVPR_2018_paper.pdf

Pseudo Mask Augmented Object Detection

In this work, we present a novel and effective framework to facilitate object detection with the instance-level segmentation information that is only supervised by bounding box annotation. Starting from the joint object detection and instance segmentation network, we propose to recursively estimate the pseudo ground-truth object masks from the instance-level object segmentation network training, and then enhance the detection network with top-down segmentation feedbacks. The pseudo ground truth mask and network parameters are optimized alternatively to mutually benefit each other. To obtain the promising pseudo masks in each iteration, we embed a graphical inference that incorporates the low-level image appearance consistency and the bounding box annotations to refine the segmentation masks predicted by the segmentation network. Our approach progressively improves the object detection performance by incorporating the detailed pixel-wise information learned from the weakly-supervised segmentation network. Extensive evaluation on the detection task in PASCAL VOC 2007 and 2012 verifies that the proposed approach is effective.

```
InProceedings{Zhao_2018_CVPR,
author = {Zhao, Xiangyun and Liang, Shuang and Wei, Yichen},
title = {Pseudo Mask Augmented Object Detection},
booktitle = {Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2018}
}
```

https://openaccess.thecvf.com/content_cvpr_2018/papers/Zhao_Pseudo_Mask_Augmented_CVPR_2018_paper.pdf

2018 CVPR
SALT LAKE CITY • JUNE 18-22

Action Sets: Weakly Supervised Action Segmentation Without Ordering Constraints

Action detection and temporal segmentation of actions in videos are topics of increasing interest. While fully supervised systems have gained much attention lately, full annotation of each action within the video is costly and impractical for large amounts of video data. Thus, weakly supervised action detection and temporal segmentation methods are of great importance. While most works in this area assume an ordered sequence of occurring actions to be given, our approach only uses a set of actions. Such action sets provide much less supervision since neither action ordering nor the number of action occurrences are known. In exchange, they can be easily obtained, for instance, from meta-tags, while ordered sequences still require human annotation. We introduce a system that automatically learns to temporally segment and label actions in a video, where the only supervision that is used are action sets. An evaluation on three datasets shows that our method still achieves good results although the amount of supervision is significantly smaller than for other related methods.

```
InProceedings{Richard_2018_CVPR,
author = {Richard, Alexander and Kuehne, Hilde and Gall, Juergen},
title = {Action Sets: Weakly Supervised Action Segmentation Without Ordering Constraints},
booktitle = {Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {June},
year = {2018}
}
```

https://openaccess.thecvf.com/content_cvpr_2018/papers/Richard_Action_Sets_Weakly_CVPR_2018_paper.pdf

Weakly- and Semi-Supervised Panoptic Segmentation

We present a weakly supervised model that jointly performs both semantic- and instance-segmentation -- a particularly relevant problem given the substantial cost of obtaining pixel-perfect annotation for these tasks. In contrast to many popular instance segmentation approaches based on object detectors, our method does not predict any overlapping instances. Moreover, we are able to segment both ``thing'' and ``stuff'' classes, and thus explain all the pixels in the image. ``Thing'' classes are weakly-supervised with bounding boxes, and ``stuff'' with image-level tags. We obtain state-of-the-art results on Pascal VOC, for both full and weak supervision (which achieves about 95% of fully-supervised performance). Furthermore, we present the first weakly-supervised results on Cityscapes for both semantic- and instance-segmentation. Finally, we use our weakly supervised framework to analyse the relationship between annotation quality and predictive performance, which is of interest to dataset creators.

https://www.ecva.net/papers/eccv_2018/papers_ECCV/papers/Anurag_Arnab_Weakly-_and_Semi-Supervised_ECCV_2018_paper.pdf

Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation

Effectively bridging between image level keyword annotations and corresponding image pixels is one of the main challenges in weakly supervised semantic segmentation. In this paper, we use an instance-level salient object detector to automatically generate salient instances (candidate objects) for training images. Using similarity features extracted from each salient instance in the whole training set, we build a similarity graph, then use a graph partitioning algorithm to separate it into multiple subgraphs, each of which is associated with a single keyword (tag). Our graph-partitioning-based clustering algorithm allows us to consider the relationships between all salient instances in the training set as well as the information within them. We further show that with the help of attention information, our clustering algorithm is able to correct certain wrong assignments, leading to more accurate results. The proposed framework is general, and any state-of-the-art fully-supervised network structure can be incorporated to learn the segmentation network. When working with DeepLab for semantic segmentation, our method outperforms state-of-the-art weakly supervised alternatives by a large margin, achieving 65.6% mIoU on the PASCAL VOC 2012 dataset. We also combine our method with Mask R-CNN for instance segmentation, and demonstrated for the first time the ability of weakly supervised instance segmentation using only keyword annotations.

https://www.ecva.net/papers/eccv_2018/papers_ECCV/papers/Ruochen_Fan_Associating_Inter-Image_Salient_ECCV_2018_paper.pdf

TS2C: Tight Box Mining with Surrounding Segmentation Context for Weakly Supervised Object Detection

This work provides a simple approach to discover tight object bounding boxes with only image-level supervision, called Tight box mining with Surrounding Segmentation Context (TS2C). We observe that object candidates mined through current multiple instance learning methods are usually trapped to discriminative object parts, rather than the entire object. TS2C leverages surrounding segmentation context derived from weakly-supervised segmentation to suppress such low-quality distracting candidates and boost the high-quality ones. Specifically, TS2C is developed based on two key properties of desirable bounding boxes: 1) high purity, meaning most pixels in the box are with high object response, and 2) high completeness, meaning the box covers high object response pixels comprehensively. With such novel and computable criteria, more tight candidates can be discovered for learning a better object detector. With TS2C, we obtain 48.0% and 44.4% mAP scores on VOC 2007 and 2012 benchmarks, which are the new state-of-the-arts.

https://www.ecva.net/papers/eccv_2018/papers_ECCV/papers/Yunchao_Wei_TS2C_Tight_Box_ECCV_2018_paper.pdf

Learning to Segment via Cut-and-Paste

This paper presents a weakly-supervised approach to object instance segmentation. Starting with known or predicted object bounding boxes, we learn object masks by playing a game of cut-and-paste in an adversarial learning setup. A mask generator takes a detection box and Faster R-CNN features, and constructs a segmentation mask that is used to cut-and-paste the object into a new image location. The discriminator tries to distinguish between real objects, and those cut and pasted via the generator, giving a learning signal that leads to improved object masks. We verify our method experimentally using Cityscapes, COCO, and aerial image datasets, learning to segment objects without ever having seen a mask in training. Our method exceeds the performance of existing weakly supervised methods, without requiring hand-tuned segment proposals, and reaches 90% of supervised performance.

https://www.ecva.net/papers/eccv_2018/papers_ECCV/papers/Tal_Remez_Learning_to_Segment_ECCV_2018_paper.pdf

Simple Does It: Weakly Supervised Instance and Semantic Segmentation

Semantic labelling and instance segmentation are two tasks that require particularly costly annotations. Starting from weak supervision in the form of bounding box detection annotations, we propose a new approach that does not require modification of the segmentation training procedure. We show that when carefully designing the input labels from given bounding boxes, even a single round of training is enough to improve over previously reported weakly supervised results. Overall, our weak supervision approach reaches 95% of the quality of the fully supervised model, both for semantic labelling and instance segmentation.

```
InProceedings{Khoreva_2017_CVPR,
author = {Khoreva, Anna and Benenson, Rodrigo and Hosang, Jan and Hein, Matthias and Schiele, Bernt},
title = {Simple Does It: Weakly Supervised Instance and Semantic Segmentation},
booktitle = {Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {July},
year = {2017}
}
```

https://openaccess.thecvf.com/content_cvpr_2017/papers/Khoreva_Simple_Does_It_CVPR_2017_paper.pdf

Weakly Supervised Cascaded Convolutional Networks

Object detection is a challenging task in visual understanding domain, and even more so if the supervision is to be weak. Recently, few efforts to handle the task without expensive human annotations is established by promising deep neural network. A new architecture of cascaded networks is proposed to learn a convolutional neural network (CNN) under such conditions. We introduce two such architectures, with either two cascade stages or three which are trained in an end-to-end pipeline. The first stage of both architectures extracts best candidate of class specific region proposals by training a fully convolutional network. In the case of the three stage architecture, the middle stage provides object segmentation, using the output of the activation maps of first stage. The final stage of both architectures is a part of a convolutional neural network that performs multiple instance learning on proposals extracted in the previous stage(s). Our experiments on the PASCAL VOC 2007, 2010, 2012 and large scale object datasets, ILSVRC 2013, 2014 datasets show improvements in the areas of weakly-supervised object detection, classification and localization.

```
InProceedings{Diba_2017_CVPR,
author = {Diba, Ali and Sharma, Vivek and Pazandeh, Ali and Pirsiavash, Hamed and Van Gool, Luc},
title = {Weakly Supervised Cascaded Convolutional Networks},
booktitle = {Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
month = {July},
year = {2017}
}
```

https://openaccess.thecvf.com/content_cvpr_2017/papers/Diba_Weakly_Supervised_Cascaded_CVPR_2017_paper.pdf

Weakly Supervised Object Localization Using Things and Stuff Transfer

We propose to help weakly supervised object localization for classes where location annotations are not available, by transferring things and stuff knowledge from a source set with available annotations. The source and target classes might share similar appearance (e.g. bear fur is similar to cat fur) or appear against similar background (e.g. horse and sheep appear against grass). To exploit this, we acquire three types of knowledge from the source set: a segmentation model trained on both thing and stuff classes; similarity relations between target and source classes; and co-occurrence relations between thing and stuff classes in the source. The segmentation model is used to generate thing and stuff segmentation maps on a target image, while the class similarity and co-occurrence knowledge help refining them. We then incorporate these maps as new cues into a multiple instance learning framework (MIL), propagating the transferred knowledge from the pixel level to the object proposal level. In extensive experiments, we conduct our transfer from the PASCAL Context dataset (source) to the ILSVRC, COCO and PASCAL VOC 2007 datasets (targets). We evaluate our transfer across widely different thing classes, including some that are not similar in appearance, but appear against similar background. The results demonstrate significant improvement over standard MIL, and we outperform the state-of-the-art in the transfer setting.

```
InProceedings{Shi_2017_ICCV,
author = {Shi, Miaojing and Caesar, Holger and Ferrari, Vittorio},
title = {Weakly Supervised Object Localization Using Things and Stuff Transfer},
booktitle = {Proceedings of the IEEE International Conference on Computer Vision (ICCV)},
month = {Oct},
year = {2017}
}
```

https://openaccess.thecvf.com/content_ICCV_2017/papers/Shi_Weakly_Supervised_Object_ICCV_2017_paper.pdf