Anisotropic Diffusion-Based Enhancement of Scene Segmentation with Instance Labels

Ioannis Kleitsiotis^{1,2}, Ioannis Mariolis¹, Dimitrios Giakoumis¹, Spiridon Likiothanassis^{1,2}, and Dimitrios Tzovaras¹

¹ Centre for Research and Technology-Hellas, Information Technologies Institute (CERTH / ITI), Thessaloniki, Greece ioklei@iti.gr

² Large Scale Machine Learning and Cloud Data Engineering Lab, Computer Engineering and Informatics, University of Patras, Patras, Greece

Abstract. Many visual scene understanding applications, especially in visual servoing settings, may require high quality object mask predictions for the accurate undertaking of various robotic tasks. In this work we investigate a setting where separate instance labels for all objects under view are required, but the available instance segmentation methods produce object masks inferior to a semantic segmentation algorithm. Motivated by the need to add instance label information to the higher fidelity semantic segmentation output, we propose an anisotropic label diffusion algorithm that propagates instance labels predicted by an instance segmentation algorithm inside the semantic segmentation masks. Our method leverages local topological and color information to propagate the instance labels, and is guaranteed to preserve the semantic segmentation mask. We evaluate our method on a challenging grape bunch detection dataset, and report experimental results that showcase the applicability of our method.

Keywords: Anisotropic · Label diffusion · Label propagation · Semantic segmentation · Instance segmentation.

1 Introduction

Vision-based understanding of complex natural scenes, e.g in agricultural settings, can provide invaluable information for in-field robotic applications. With the success of deep learning frameworks for end-to-end object detection during the past years, new opportunities arise for more complete and robust field scene understanding. However, the biological entities which are needed to be considered for robotic applications are often highly complex. Especially in vineyards, the unstructured geometrical configuration of the vine plant parts which are more important for robotic applications, e.g grape bunches, trunks, cordons and shoots, lead to significant challenges for the state of the art object detection algorithms.

We are concerned with the instance segmentation problem, where the masks and instance labels for each countable object on the scene must be predicted.

The caveat that motivates our method is that a semantic segmentation algorithm might predict object masks of higher quality than the corresponding masks from an instance segmentation algorithm. Therefore, our goal is the addition of instance labels to the higher fidelity semantic segmentation algorithm, with the guarantee that the original segmentation mask will be preserved. We formulate the 2D instance label enhancement problem of semantic segmentation as an anisotropic label diffusion process, taking inspiration both from the 3D label diffusion process presented in [25] and the generalization of label diffusion to the anisotropic case [13]. More specifically, our method leverages local pixel information to propagate all predictions from an instance segmentation algorithm intersecting with our semantic segmentation predictions, towards the remaining semantic segmentation predictions.

To the best of our knowledge, this is the first work to address the incorporation of instance labels in a deep semantic segmentation framework through the anisotropic diffusion of instance labels generated by a lower fidelity instance segmentation algorithm guided by topological and color information. We experimentally evaluate our results in a vineyard grape bunch setting using two representative state-of-the-art methods, PSPNet [31] for semantic segmentation and Mask R-CNN [11] for instance segmentation. We found PSPNet to achieve higher mask quality than Mask R-CNN in this setting, and thus, we preserved PSPNet predictions while simultaneously enhancing them with instance label information.

2 Related Work

Although several works combine semantic and instance segmentation for the task of panoptic segmentation [10, 19, 28, 26, 7], first described in [14], some works explicitly aim at mask improvement through the combination of the two tasks. The authors of [8] add an Atrous convolution segmentation head to Mask R-CNN to refine the predicted segmentation masks. In [24], a semantic segmentation U-Net [21] head is attached in Mask Scoring R-CNN [12] to facilitate mask prediction performance. The authors of [18] utilize a Bayesian setting to improve [16] with semantic segmentation masks from [23]. The authors of [9] present a novel Instance Mask Projection neural module that enhances segmentation mask prediction by projecting Mask R-CNN masks to the features of the semantic segmentation module. In [5] the authors improve the segmentation results of UNet in medical images first by estimating bounding boxes of anatomical structures with a connected components analysis, and afterwards by combining Mask R-CNN and UNet through a bounding box tracking-based approach. The tracking unfolds across 2D slices of the original Computed Tomography 3D volumes. The authors of [17] enforce consistency between semantic and instance segmentation masks prediction through a specialized loss construction. In [30], the authors add a semantic segmentation head in Mask R-CNN to uncover fine details in a crack detection setting. All the aforementioned works combine semantic and instance segmentation in a fusion setting, where it is assumed that

the combination of the two tasks will facilitate the overall mask predictions. We on the other hand aim to preserve only the semantic segmentation masks, while simultaneously adding instance label information.

Two works that perform instance segmentation by applying the watershed transform on a semantic segmentation mask, and thus, like our method, retain the initial semantic segmentation mask quality, are [4] and [29]. In the first paper, the authors train a neural network to learn the energy basins of the watershed transform with an intermediate step of distance transform learning, and apply thresholds at predefined energy values to extract the object instances. The authors of the second work directly calculate the euclidean distance transform of the semantic segmentation output of a UNet [21] (based on the implementation found in [1]), and apply the watershed transform on it with markers provided by the centers of bounding boxes estimated by a modified Region Proposal Network (from Faster R-CNN [20]). Although these works preserve the semantic segmentation mask, the instance segmentation quality depends on the watershed transform performance, while our work aims at a modular design that permits the selection of an appropriate instance segmentation algorithm and offers controllable quality in instance label detection.

The work closest to ours is [3], where the authors combine the output of a semantic segmentation network with bounding boxes predicted by an object detector to perform instance segmentation. Their method retains semantic segmentation masks and assigns to the regions not covered by bounding boxes instance labels based on the mean field approximation of a CRF [15], which is a non-local [6] process. On the other hand, our method is mainly concerned with the utilization of instance mask predictions that have considerable overlap with the ground truth, i.e each visible part of an object instance is covered by a coarse instance mask, and therefore does not aim to assign instance information in a global-scale, but smoothly propagate coarse predictions towards unlabeled pixels. The addition of anisotropic guidance to the diffusion process guarantees the prevention of possible "leakage" of instance label information between different instances.

3 Methodology

Let L be the set of object instances occurring in an RGB image I of width W and height H , M_I^l be the estimated mask from the instance segmentation algorithm for label $l \in L$, M_S the mask predicted by the semantic segmentation algorithm, and M_I the logical OR for all Mask R-CNN masks, i.e $\bigvee_{l=0}^{|L|} M_I^l$. For each pixel $p_i \in M_S$ we find the set of its K -nearest neighbors based on the euclidean distance between the pixels, denoted as $KNN(p_i)$. We search for nearest neighbors only inside M_S , i.e $p_j \in KNN(p_i) \Rightarrow p_j \in M_S$.

We plan to propagate the instance labels predicted by the instance segmentation algorithm via anisotropic diffusion, towards the unlabeled pixels belonging to M_S but not to M_I . To this end, we first define the set $z = \{z^l\}$ of label vectors $z^l \in \{0, 1\}^{WH}$ with $l \in L$. We set the vector element $z^l(i)$ corresponding

Fig. 1. The overall pipeline of our method: The intersection of the semantic segmentation and the instance segmentation outputs are combined to provide the initial instance labels, which are afterwards anisotropically diffused towards the remaining unlabeled pixels of the Semantic Segmentation output. The $\cap$ symbol signifies mask intersection.

to pixel p_i to be equal to 1 if $p_i \in M_I^j$, and zero otherwise. Then, we define the weighted graph $G = (P, E, W)$ with the set of nodes P equal to the mask pixels M_S , edges $E \subset P \times P$ connecting the nodes and weights W assigned to every edge. Each z_i^l defines a function connecting the i -th pixel/graph node p_i with the value z_i^l indicating whether p_i belongs to instance label l or not. Let $\|\cdot\|_2$ denote the euclidean norm and $I(p_i)$ the RGB color value of pixel p_i . Then, the anisotropic graph laplacian $L^D(i)$ on graph node i is defined as [13]:

$$L^D(i) = \left(\frac{1}{d_i} \sum_{j=1}^n w_{ij} q_{ij} \right) z_i^l - \frac{1}{d_i} \sum_{j=1}^n w_{ij} q_{ij} z_j^l \quad (1)$$

where w_{ij} is the edge weight connecting nodes i and j :

$$w_{ij} = \exp\left(\frac{\|p_i - p_j\|_2^2}{\sigma_d^2}\right) \cdot \exp\left(\frac{\|I(p_i) - I(p_j)\|_2^2}{\sigma_c^2}\right) \quad (2)$$

d_i is the degree of node i :

$$d_i = \sum_{j=1}^n w_{ij} \quad (3)$$

and q_{ij} is equal to:

$$q_{ij} = \exp\left(-\frac{w_{ij}(z_i^l - z_j^l)^2}{\sigma_a^2}\right) \quad (4)$$

where σ_d , σ_c and σ_a are the scale hyperparameters of the exponents.

Assuming a unit discretization interval, the Euler approximation for anisotropic diffusion on G per instance label class is given by the following iterative update from step t to step $t + 1$:

$$z_{t+1}^l \leftarrow z_t^l - L^D z_t^l \quad (5)$$

The total number of iterations is a hyperparameter - following [13], we first apply isotropic diffusion ($q_{ij} = 1$) for a small number of steps to decrease the number

of iterations required for convergence. After the final iteration is completed for every instance l , we select the instance label l_{p_i} of every initially unlabeled pixel p_i to be equal to:

$$l_{p_i} = \arg \max_l (z^l) \quad (6)$$

A schematic overview of our method can be seen in Fig. 1.

4 Experimental Evaluation

Our method was tested on the publicly available Embrapa WGISD [22] dataset. Embrapa WGISD contains 137 RGB images with manually annotated instance segmentation masks of grape bunches. We follow the 88-22-27 training-validation-test images split used by the authors, with the same image resizing and augmentation strategies described in their paper. In our experiments, we found the Detectron2 [27] implementation of Mask R-CNN to produce superior results to the Matterport [2] implementation used in [22]. We trained both implementations of Mask R-CNN, and the implementation of PSPNet [33], [32], in an Nvidia Tesla K40m GPU, with the maximum available batch size and network backbone. The Detectron2 Mask R-CNN implementation was trained for 6000 iterations, the PSPNet implementation for 10000 iterations and the Matterport Mask R-CNN implementation for 6000 iterations. For our experiments, the instance segmentation masks from Mask R-CNN were predicted using the confidence value that maximized the (object detection) F-measure at 0.3 IoU of the validation set. In all frameworks, the default training parameters were kept.

We evaluate the mask quality of PSPNet and Mask R-CNN with total mask Grape Bunch IoU, total mask Background IoU (BG IoU) and per-pixel F-measure. We also include the F-measure metric reported in [22] for the Matterport Mask R-CNN implementation. The total scene masks for Mask R-CNN were derived by aggregating all predicted instance masks using the logical OR operator. The experimental results of Table 1 highlight the relevance of our method in this dataset, as the semantic segmentation algorithm predicts higher quality masks compared to the instance segmentation algorithm.

Table 1. Results on Embrapa WGISD Dataset

Segmentation Method	Grape Bunch IoU	BG IoU	F-measure
Matterport Mask R-CNN([22])	-	-	0.889
Matterport Mask R-CNN(Ours)	0.792	0.979	0.884
Detectron2 Mask R-CNN	0.834	0.983	0.909
PSPNet	0.858	0.985	0.924

Therefore, we apply anisotropic diffusion-based instance label addition to the test set. Based on our experimental hyperparameter search, we utilized the

8 nearest neighbors and set $\sigma_d = 2$, $\sigma_c = 500$ and $\sigma_a = 0.1$. We also used the Detectron2 Mask R-CNN implementation to provide the initial instance labelling. In Table 2 we compare our method to a naive local algorithm that for each unlabeled pixel p , uses the label majority vote of the K nearest labeled neighbors to assign to p an instance label. We refer to this baseline method as Majority Voting KNN (MV-KNN), and set the number of neighbors equal to 9. For labelled neighbors existing on Mask R-CNN mask overlaps, we select the contributing label randomly from one of the overlapping masks. We compare the methods using per-unlabeled-pixel average multi-class precision (Prec), recall (Rec) and F-measure. We found that our method produces adequate results even for approximately 100 diffusion iterations, and reaches convergence with respect to the evaluation metrics at approximately 1000 iterations.

Fig. 2. PSPNet output enhancement with grape bunch instance label information. The columns from left to right depict: the ground truth instance segmentation labels, the PSPNet semantic segmentation output, the Mask R-CNN (Detectron2) instance segmentation results, and the results of our method.

However, to have meaningful metrics, we remove the influence of wrong predictions from the semantic and instance segmentation algorithms. We remove the influence of object misclassification from Mask R-CNN by assigning to each pixel of the predicted instance segmentation masks the original ground truth instance labels - i.e, we still propagate from masks predicted by Mask R-CNN, but with instance labels given by the ground truth. Additionally, to disregard the effect of possible PSPNet mask inaccuracies, we calculate the metrics only on correct PSPNet mask estimations.

Table 2. Comparison of Our Method with the Baseline

Method	Prec	Rec	F-measure
MV-KNN	0.841	0.907	0.843
Ours	0.904	0.911	0.893

Our method outperforms the baseline in all metrics, and achieves over 6% improvement in precision. Some qualitative results that highlight the effectiveness of our method can be seen in Fig. 2. The reader should notice that in the fourth column of Fig. 2 several Mask R-CNN instance masks have been improved. Regions from grape bunch masks that were improved by preserving PSPNet mask information have been noted with black arrows in the third column. The resolution of conflicts between adjacent bunch grape masks attempting to cover the same unclaimed PSPNet mask region, for example in the outermost right grape bunches (yellow and purple) in the second row, is guided by the local topological and color information in the diffusion process. It should be noted that the ground truth coloring of grape bunch masks is different from Mask R-CNN and our method, due to the fact that Mask R-CNN usually predicts a different number of instance masks, and in a different order.

5 Conclusion

An anisotropic diffusion-based instance labelling algorithm for semantic segmentation masks has been presented. Our method utilizes local distance and color cues to anisotropically propagate instance label information predicted by an instance segmentation algorithm towards mask pixels estimated by a semantic segmentation algorithm. The evaluation of our method in the Embrapa WGSD grape bunch instance segmentation dataset showcases the preservation of semantic segmentation mask quality. Our future work will concentrate on the fusion of RGB and depth information in the diffusion process to alleviate possible ambiguities in scene understanding originating from grape bunch instance segmentation misassignments. We also plan to evaluate our method on datasets of larger size and variability.

Acknowledgment

This work has been supported by the EU Horizon 2020 funded project namely: “BACCHUS (MoBile Robotic PLATforms for ACtive InspeCtion and Harvesting in AgricUltural AreaS)” under the grant agreement with no: 871704.

References

1. Nuset: A deep learning tool for reliably separating and analyzing crowded cells. <https://github.com/yanglf1121/NuSeT>
2. Abdulla, W.: Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN (2017)
3. Arnab, A., Torr, P.: Bottom-up instance segmentation using deep higher-order crfs. *ArXiv* (2016)
4. Bai, M., Urtasun, R.: Deep watershed transform for instance segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2858–2866 (2017)
5. Bouget, D., Jørgensen, A., Kiss, G., Leira, H.O., Langø, T.: Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in CT data for lung cancer staging. *Int J Comput Assist Radiol Surg* **14**(6), 977–986 (Jun 2019)
6. Caye Daudt, R., Le Saux, B., Boulch, A., Gousseau, Y.: Guided anisotropic diffusion and iterative learning for weakly supervised change detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
7. Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
8. Costea, A., Petrovai, A., Nedevschi, S.: Fusion scheme for semantic and instance-level segmentation. 2018 21st International Conference on Intelligent Transportation Systems (ITSC) pp. 3469–3475 (2018)
9. Fu, C.Y., Berg, T.L., Berg, A.C.: Imp: Instance mask projection for high accuracy semantic segmentation of things. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
10. de Geus, D., Meletis, P., Dubbelman, G.: Panoptic segmentation with a joint semantic and instance segmentation network. *ArXiv* **abs/1809.02110** (2018)
11. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
12. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask Scoring R-CNN. In: CVPR (2019)
13. Kim, K.I., Tompkin, J., Pfister, H., Theobalt, C.: Context-guided diffusion for label propagation on graphs. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2776–2784 (2015)
14. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9396–9405 (2019)

15. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*. vol. 24. Curran Associates, Inc. (2011)
16. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation (2017)
17. Liu, D., Zhang, D., Song, Y., Huang, H., Cai, W.: Panoptic feature fusion net: A novel instance segmentation paradigm for biomedical and biological images. *IEEE Transactions on Image Processing* **30**, 2045–2059 (2021). <https://doi.org/10.1109/TIP.2021.3050668>
18. Pham, V., Ito, S., Kozakaya, T.: Biseg: Simultaneous instance segmentation and semantic segmentation with fully convolutional networks. *CoRR* **abs/1706.02135** (2017), <http://arxiv.org/abs/1706.02135>
19. Porzi, L., Rota Bulò, S., Colovic, A., Kotschieder, P.: Seamless scene segmentation. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 28. Curran Associates, Inc. (2015)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. pp. 234–241. Springer International Publishing, Cham (2015)
22. Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S.: Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Computers and Electronics in Agriculture* **170**, 105247 (2020)
23. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**, 1–1 (05 2016)
24. Tian, Y., Yang, G., Wang, Z., Li, E., Liang, Z.: Instance segmentation of apple flowers using the improved mask r-cnn model. *Biosystems Engineering* **193**, 264–278 (2020)
25. Wang, B.H., Chao, W., Wang, Y., Hariharan, B., Weinberger, K.Q., Campbell, M.: Ldls: 3-d object segmentation through label diffusion from 2-d images. *IEEE Robotics and Automation Letters* **4**(3), 2902–2909 (2019)
26. Wu, Y., Zhang, G., Gao, Y., Deng, X., Gong, K., Liang, X., Lin, L.: Bidirectional graph reasoning network for panoptic segmentation. In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 9080–9089 (2020)
27. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
28. Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: Upsnet: A unified panoptic segmentation network (2019)
29. Yang, L., Ghosh, R.P., Franklin, J.M., Chen, S., You, C., Narayan, R.R., Melcher, M.L., Liphardt, J.T.: Nuset: A deep learning tool for reliably separating and analyzing crowded cells. *PLOS Computational Biology* **16**(9), 1–20 (09 2020)
30. Zhang, Y., Chen, B., Wang, J., Li, J., Sun, X.: APLCNet: Automatic pixel-level crack detection network based on instance segmentation. *IEEE Access* **8**, 199159–199170 (2020)

31. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239 (2017)
32. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
33. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision* (2018)