Open-World Instance Segmentation: Exploiting Pseudo Ground Truth From Learned Pairwise Affinity

Supplementary material

Weiyao Wang[†], Matt Feiszli[†], Heng Wang[†], Jitendra Malik[†§], Du Tran[†]
[†] Meta AI Research [§] UC Berkeley
{weiyaowang,mdf,hengwang,trandu}@fb.com, malik@berkeley.edu

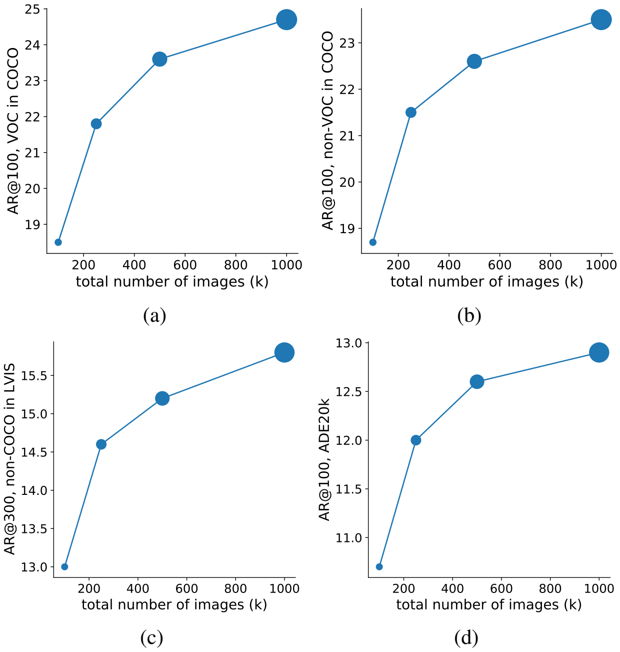
In this supplementary material, we include:

1. Experiments on OpenImages [11] demonstrating that scaling number of unlabeled training images further improves GGN (section A).
2. Effects of ImageNet [5] pre-training on open-world instance segmentation (section B)
3. Additional qualitative results of open-world segmentation in the wild on ADE20K and UVO (section C).
4. A proof of concept experiment of using GGN for closed-world class-aware instance segmentation (section D).
5. Our discussion on the limitations and future directions of the proposed method (section E).
6. Ablations on data augmentation techniques for training PA (section F)

A. Improve GGN by scaling unlabeled pixels

In section 5 of main paper, we showed how our proposed method generates pseudo-GT masks on unlabeled images, and how GGN benefited from training on unlabeled images (Table 7). Here, we further show how GGN can be further improved by scaling the number of unlabeled training images.

We increase the size of unlabeled images (e.g., 100k, 250k, 500k, 1M) sampled from OpenImagesV4 [11] and take top-3 scoring pseudo-masks per image and use them as pseudo-GT masks for training. As shown in Figure 1, increasing the number of unlabelled training images continuously improves model performances in various setups. This further demonstrates the potential of GGN in both open-world (non-VOC, non-COCO [12], ADE20k [17]) and closed-world (VOC) instance segmentation.



(a) (b) (c) (d)

Figure 1. **The effect of scaling the number of images in training GGN.** We increase the size of subset of OpenImages [11] to 100k, 250k, 500k and 1M and train GGNs with pseudo-masks generated by pairwise affinities trained on VOC masks. In all setups, scaling images keeps improving model performance.

B. ImageNet pre-training for open-world instance segmentation

In closed-world setup, ImageNet label pre-training offers limited values [8]: when training from scratch at 6x standard schedule, detectors perform on-par with 1x schedule finetuning from ImageNet label pre-training. This questioned if ImageNet label pre-training is a strong baseline to compare to (as we did in section 5.5). We argue that it is indeed a strong baseline, and that ImageNet label pre-training outperforms 6x schedule training from scratch (Table 1). This validates the value of ImageNet label pre-training for

Training strategy	LVIS	UVO	ADE20K
6x schedule from scratch	13.1	41.5	13.7
ImageNet pre-training	**14.7**	**41.8**	**14.6**

Table 1. **Different from common wisdom in closed-world instance segmentation, ImageNet pre-training outperforms long training schedule from random initialization in open-world**. We verify this with Mask R-CNN trained/ finetuned on COCO and evaluate on Non-COCO categories in LVIS [6], UVO [16] and ADE20K [17]

open-world instance segmentation, making it a proper baseline to compare with.

C. Qualitative results in the wild

We provide additional visualizations to compare GGN and baseline Mask R-CNN on ADE20k and UVO [16] (Figure 2 and Figure 3). Both models are trained with masks from 80 COCO categories, with GGN enhanced by pseudo-masks on COCO images. We show that GGN can recall more true positive segments than baseline, including novel objects, severely occluded objects and stuff.

D. Does generic grouping help closed-world segmentation?

In the previous experiments, we showed that GGN is useful for instance segmentation in the open-world (a.k.a class-aware instance segmentation). One may wonder if GGN is also useful for closed-world segmentation. In order to answer the question, we conduct the following proof-of-concept experiment. We adopt the standard Mask R-CNN [9] by replacing its RPN branch with our GGN. We note that our GGN also outputs bounding boxes and masks thus can completely replace RPN. In our experiment, GGN is pretrained with pseudo-GT in a class-agnostic and fixed (no fine-tuning or refinement) during class-aware training and evaluation. This means, during class-aware training and evaluation, only the recognition head is trained. We hypothesis the GGN can be competitive with RPN, even with a closed-world, class-aware setup. We name this modified architecture as *Two-Tower* to reflex the recognition and grouping branches. The grouping branch, GGN, is trained on only VOC-category masks. We compare this Two-Tower architecture with Mask R-CNN which is trained end-to-end in limited data domain: using only 10% of COCO images on all classes. Whereas Mask R-CNN is trained on grouping from all categories, the two-tower grouping module only leverages VOC masks and generated pseudo masks. Results are presented in Table 2.

E. Limitations and future directions

We present GGN that combines bottom-up grouping and top-down training for open-world instance segmentation. The framework has shown significant gains and achieves the new state-of-the-art results on multiple benchmarks. In this section, we discuss the limitations of the approach, which also inform future directions to tackle.

Training length	Method	mAP	mAR
short	Mask R-CNN	10.6	36.2
	Two-Tower	**12.7**	**36.5**
normal	Mask R-CNN	**15.4**	**40.0**
	Two-Tower	13.5	36.5

Table 2. **Proof of concept on Two-Tower model for grouping and recognition** Mask R-CNN is trained on all 80 COCO categories. GGN, as the grouping module, is only trained on 20 VOC classes. The recognition module does not alter the mask predictions of the grouping module, and is trained with 80 COCO categories for classification. Two-tower is competitive in both short and normal training schedules.

Objectness. In GGN, we used WT+UCM [2] to group pixels into segments leveraging learned pixel pairwise affinities. However, WT+UCM has certain limitations: it has no notion of objectness, and therefore constructs pixel groups of “part” of an object. It is important to find novel methods to select good masks from all proposed pseudo-masks leveraging certain objectness prior, which can be learned [10] or hand-crafted [3].

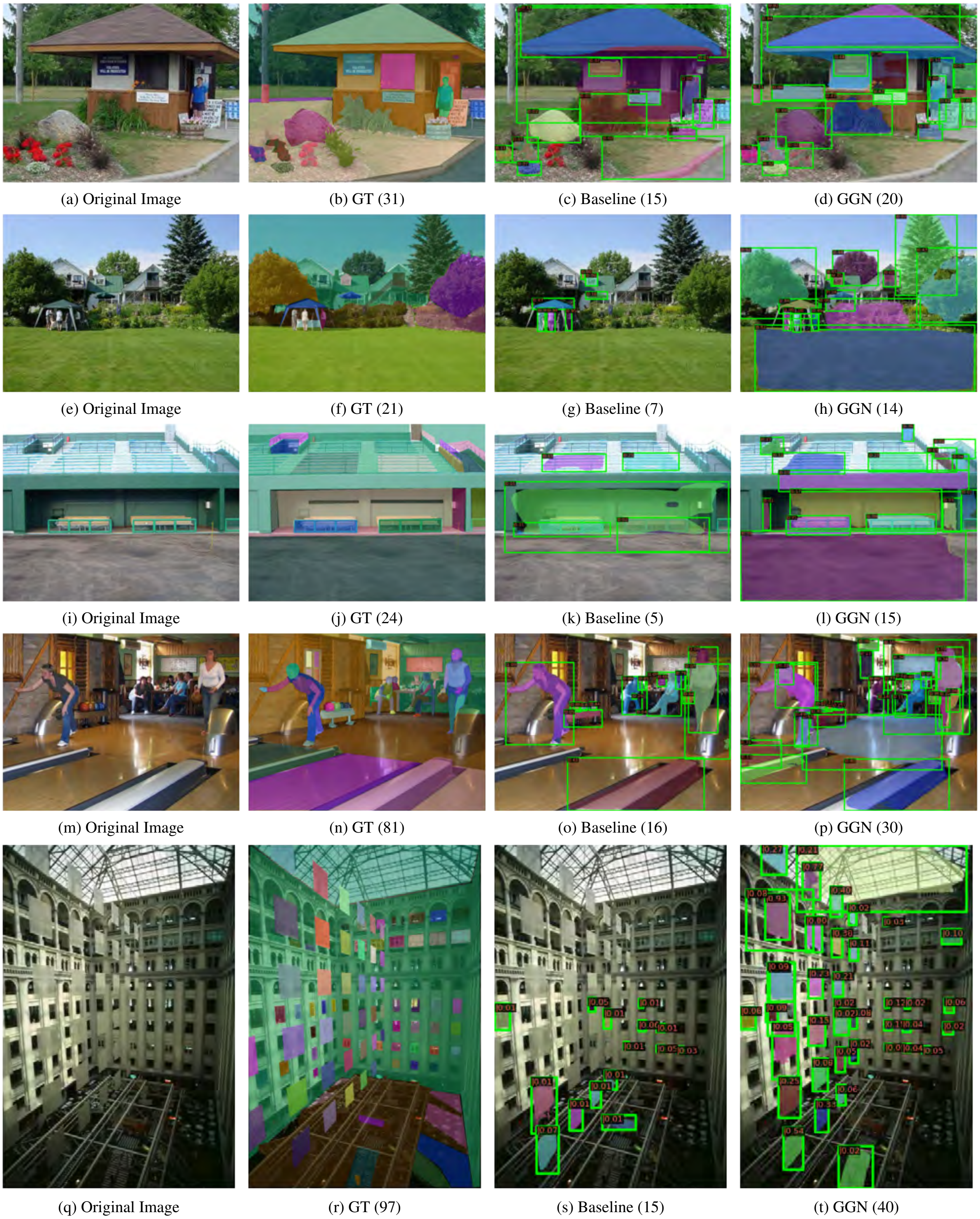
Hierarchy of groups. When we select pseudo-GT masks generated from pairwise affinities, we ignore the natural hierarchical structure of the groups generated by UCM. It is worth understanding if enforcing grouping hierarchies can further improve the supervision signals.

Grouping as pretext task. Existing frameworks, such as Mask R-CNN, leverage recognition as pre-training for grouping (e.g., by pre-training on ImageNet). In this paper, we have demonstrated the value of training on unlabeled data to form grouping. A extension of this work should study how learning to group can potentially benefit recognition ability.

F. Data augmentation for learning PA

While data augmentation is well-explored in learning object proposals or masks [15, 18], it is not well-studied in the context of pairwise affinities or similar representation such as semantic edges [1, 13]. Different from bounding boxes or masks, pairwise affinities are local features and can be very sensitive to both pixel-level and spatial-level transforms.

Many data augmentation has a positive effect on pairwise affinities: multi-scaling is the strongest augmentation among all. Besides, CLAHE [14] and Hue-Saturation value jittering provide strong pixel-level augmentation. Not all augmentation helps to learn pairwise affinities: among 20 types of augmentation experimented, more than half hurts the performance of pairwise affinities (Fig. 4). For instance, different from the findings in contrastive learning [4, 7], all kinds of blurring hurts the performance of pairwise affini-



(a) Original Image (b) GT (31) (c) Baseline (15) (d) GGN (20)

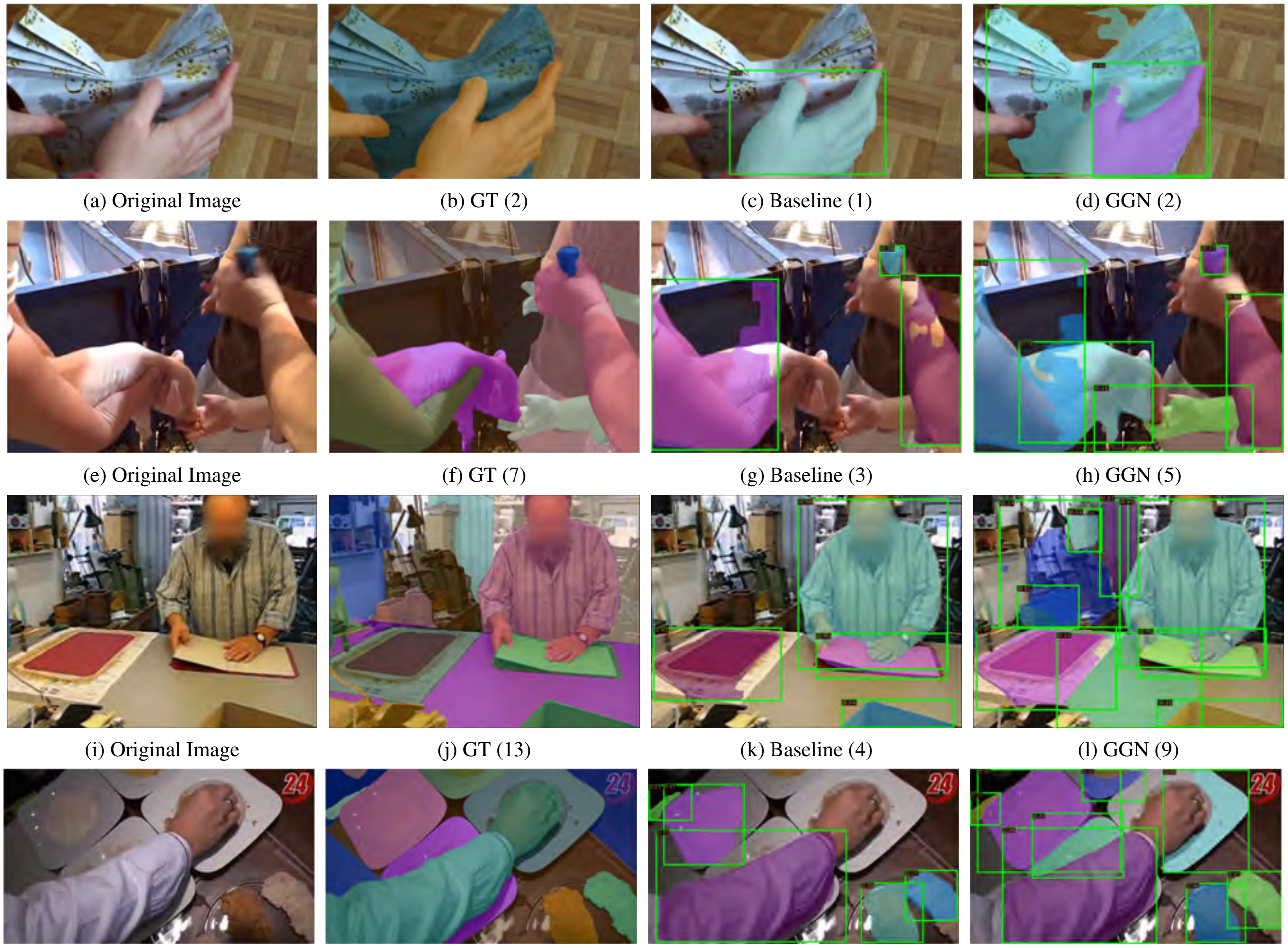
(e) Original Image (f) GT (21) (g) Baseline (7) (h) GGN (14)

(i) Original Image (j) GT (24) (k) Baseline (5) (l) GGN (15)

(m) Original Image (n) GT (81) (o) Baseline (16) (p) GGN (30)

(q) Original Image (r) GT (97) (s) Baseline (15) (t) GGN (40)

Figure 2. **Visualization of GGN compared to baseline on ADE20k.** We take top-100 scoring predictions for each of the methods. GGN detects significantly more true positive segments compared to baseline, including novel objects and stuff. Number in bracket represents number of retrieved segments.



(a) Original Image (b) GT (2) (c) Baseline (1) (d) GGN (2)

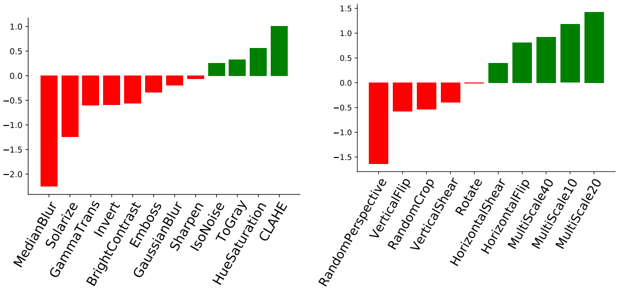
(e) Original Image (f) GT (7) (g) Baseline (3) (h) GGN (5)

(i) Original Image (j) GT (13) (k) Baseline (4) (l) GGN (9)

(m) Original Image (n) GT (11) (o) Baseline (5) (p) GGN (8)

Figure 3. **Visualization of GGN compared to baseline on UVO.** We take top-100 scoring predictions for each of the methods. GGN detects significantly more true positive segments compared to baseline, including novel objects and stuff. Number in bracket represents number of retrieved segments.

ties. Pairwise affinities predict local relationship, which becomes uncertain with blurred images. In addition, orientation matters. While horizontal flipping and shearing contribute positively to learning pairwise affinities, vertical operators of the same kinds hurt the performance. We visualize a few augmentation in Figure 5.



(a) (b)

Figure 4. **The effects of different data augmentations on learning pairwise affinities.** The performance is evaluated by UCM masks generated by the pairwise affinities trained under different augmentations. Performance is represented as gain (loss) in AR100 compared to without augmentation.

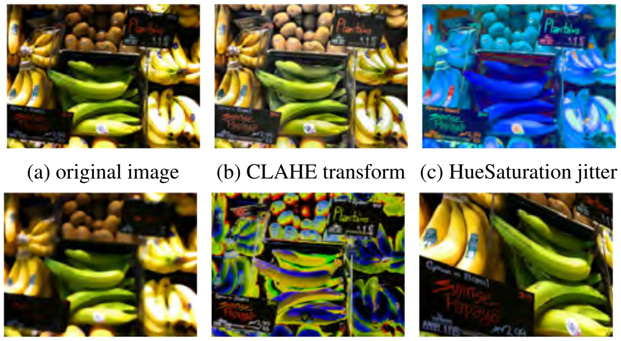
References

[1] David Acuna, Amlan Kar, and Sanja Fidler. Devil is in the edges: Learning semantic boundaries from noisy annotations. In *CVPR*, 2019. 2

[2] Pablo Arbelaez. Boundary extraction in natural images using ultrametric contour maps. In *CVPR Workshops*, 2006. 2

[3] Pablo Andrés Arbeláez, Jordi Pont-Tuset, Jonathan T. Barron, Ferran Marqués, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2



(a) original image (b) CLAHE transform (c) HueSaturation jitter

(d) Median blur (e) Solarize (f) Random perspective

Figure 5. **Visualization of data augmentation.** Top row includes original image and two strong pixel-level augmentation. Bottom row contains three augmentation types that hurt the performance.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[6] A. Gupta, P. Dollár, and R. Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5351–5359, 2019. 2

[7] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2

[8] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4917–4926, 2019. 1

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017. 2

[10] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *CoRR*, abs/2108.06753, 2021. 2

[11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 1

[12] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[13] Yun Liu, Ming-Ming Cheng, Jiawang Bian, Le Zhang, Peng-Tao Jiang, and Yang Cao. Semantic edge detection with diverse deep supervision. *ArXiv*, abs/1804.02864, 2018. 2

[14] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987. 2

[15] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. In *arXiv:2005.04757*, 2020. 2

[16] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 2

[17] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 1, 2

[18] Barret Zoph, Ekin Dogus Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. Learning data augmentation strategies for object detection. In *ECCV*, 2020. 2