Reviewer #1 *Differences to Zhang et al., 2018 (USD)* : Important point! USD introduces an explicit noise modeling to capture the noise and generate noise-free pseudo labels [L38-50]. This method produces blurry saliency predictions. On the contrary, we refine the pseudo-labels individually to preserve diversity and enforce inter-image consistency before fusing all the pseudo-labels, which is crucial for producing sharp and fine details of the salient objects.

Computational efficiency: The proposed framework needs extra computation for refining handcrafted methods in isolation. However, the saliency prediction network converges faster than USD once the refined labels are available.

List the saliency detection methods: We use the handcrafted methods MC, HS, DSR, RBD [L197-198] in our work.

Reviewer #1 & #2 *Value to CV/ML-research or technical contribution*: Existing unsupervised techniques combine and reuse the handcrafted methods by directly adapting the noisy pseudo labels. We are the first to refine the labels from these methods individually in isolation and incrementally improve them with self-supervision via historical model averaging. This improves the results substantially, being on-par even with supervised learning while reducing the manual labeling effort.

Reviewer #2 *Modify style and acronyms*: To mitigate the confusion and improve the flow, we will place the related works right after the introduction and incorporate the suggestions concerning the acronyms in our final version. Thanks for pointing it out and helping us to improve the readability.

Evaluation on ECSSD, DUT, SED2: Yes, we follow the exact evaluation procedure of Zhang et al. 2018 and use the model trained on MSRA-B to evaluate on these datasets. This is common practice for object saliency prediction.

Compare to Chen et al. 2018 (DRN): DRN was developed for n-class semantic segmentation and was evaluated only on this task. Inspired by its impressive results, we used DRN as backbone of our framework to make binary object saliency predictions. We cannot compare our results on saliencies to semantic segmentation results from Chen et al.

No evaluation on Pascal segmentation: This dataset has non-binary labels (no binary ground-truth labels for object saliency prediction) which impede the computation of the F-score measure.

Is F-measure pixel-wise and how does it preserve inter-images consistency: The F-measure is computed across the pixels in the image and not pixel-wise. Thanks for pointing it out. We will remove the word "pixel-wise" from L102 to avoid the confusion and reformulate this part. The pseudo-label generation network trained on entire dataset enforces the inter-images consistency. The handcrafted methods do not leverage the features from images, whereas the deep network learns to produce consistent output maps from the training images, as shown in Figure 2.

'No-CRF' and 'no self-supervision' in tables: No-CRF implies that we do not apply a CRF to the final outputs of the network. This variant reduces the inference time for time-critical applications. No self-supervision indicates leaving out "incremental refining via self-supervision" (Fig.4c) from the framework.

Changes to DRN and ResNet: The last layer of DRN produces multiple class outputs for semantic segmentation. We modified this last layer to yield binary images, as needed for our saliency prediction framework, and trained the entire network, including the last layer. Analogous changes are applied to ResNet.

Why fix the no. of training iterations to 25: We observed that network training reaches a coarse convergence on an error plateau when combined with a small learning rate. Optimizing this hyper-parameter might lead to better performance.

Other forms of regularization: we investigated other techniques such as adversarial training, auxiliary losses with inpainting or reconstruction. We found that minor improvement does not justify the added complexity of our system.

Reviewer #3 *Mention connection to crowdsourcing*: Great suggestion! We will mention it in our final version.

Avoid cumulative mistakes: Given the labels' diversity among different handcrafted methods, the accumulated mistakes are typically outnumbered in the final fusion step. It is unlikely that multiple methods make the same mistake.

Under what condition noisy can label refinement be helpful for better results: The noisy labels provide weak supervision, which misleads the learning process and thus affects the network generalization. Refinement of the noisy labels improve supervisory signal (similar to fully supervised setting), stabilizes training and enhances generalization of the network.

Influence of the number of handcrafted methods on the final label quality: The diversity of the pseudo-labels created by different handcrafted methods is essential and actually more important than their absolute number. In Table 2, we compare the performance of the full model to the saliency prediction network trained using labels attained from only a single handcrafted method. The difference shows the importance of pseudo-labels from diverse methods.

Failures cases: We observe large overlapping with the traditional supervised learning methods in this regard. The failures comprise corner cases like small objects and shadows. We will add failure cases in the final version.

Do refined pseudo-labels need to be fused selectively? Our framework shows that selective fusion is not necessary. However, a clever fusion scheme may potentially further improve the system's performance.

Fig. 5: The curves (b & d) show the quality of MVA pseudo labels (the similarity of labels w.r.t. ground-truth) of every handcrafted method at every step in our pipeline. The curves (a & c) show the differences in quality of saliency map predictions obtained with the network trained on MVA pseudo labels retrieved at different steps in the pipeline.