A Novel Benchmark for Refinement of Noisy Localization Labels in Autolabeled Datasets for Object Detection — Supplementary Material

Andreas Bär[1] * Jonas Uhrig[2 3] * Jeethesh Pai Umesh[1] † * Marius Cordts[2] Tim Fingscheidt[1]

[1]Technische Universität Braunschweig, Germany [2]Mercedes Benz AG, Germany [3]University of Freiburg, Germany

*Equal contribution †Work done during an internship at Mercedes Benz AG, Germany

{andreas.baer, j.umesh, t.fingscheidt}@tu-bs.de {jonas.uhrig, marius.cordts}@mercedes-benz.com

1. More Details About the Losses

We introduce the smooth L_1 loss [1] and the generalized intersection-over-union (gIoU) loss [2].

The smooth L_1 loss is defined as

$$J^{\text{smooth}} = \begin{cases} \Delta - \frac{\theta}{2}, & \text{if } \Delta \geq \theta, \\ \frac{1}{2\cdot\theta}\Delta, & \text{otherwise}, \end{cases} \tag{1}$$

with $\Delta = |\boldsymbol{y} - \overline{\boldsymbol{y}}|$ or $\Delta = |\boldsymbol{y} - \widetilde{\boldsymbol{y}}_k|$ depending on the `LLRN` type and iteration k. We use $\theta = 1$.

Further, let $\mathcal{Y}_{\overline{o}}$, $\widetilde{\mathcal{Y}}_{\overline{o}}$, and $\overline{\mathcal{Y}}_{\overline{o}}$ contain all pixel positions $(h, w) \in \mathcal{I}$ that are enclosed by the object localization represented as $\boldsymbol{y}_{\overline{o}}$, $\widetilde{\boldsymbol{y}}_{\overline{o}}$, and $\overline{\boldsymbol{y}}_{\overline{o}}$, respectively. In addition, let $\mathring{\boldsymbol{y}}_{\overline{o}} = \big((\mathring{h}_{\overline{o}}^{(\mathrm{L})}, \mathring{w}_{\overline{o}}^{(\mathrm{L})}), (\mathring{h}_{\overline{o}}^{(\mathrm{R})}, \mathring{w}_{\overline{o}}^{(\mathrm{R})})\big)$, with top-left corner $\big(\mathring{h}_{\overline{o}}^{(\mathrm{L})}, \mathring{w}_{\overline{o}}^{(\mathrm{L})}\big)$ and bottom-right corner $\big(\mathring{h}_{\overline{o}}^{(\mathrm{R})}, \mathring{w}_{\overline{o}}^{(\mathrm{R})}\big)$ be the smallest possible enclosing box of two object localizations referring to an object with identifier $\overline{o}$, e.g., $\boldsymbol{y}_{\overline{o}}$ and $\overline{\boldsymbol{y}}_{\overline{o}}$. The top-left (L) and bottom-right (R) corners of $\mathring{\boldsymbol{y}}_{\overline{o}}$ are then computed by

$$\begin{aligned} \big(\mathring{h}_{\overline{o}}^{(\mathrm{L})}, \mathring{w}_{\overline{o}}^{(\mathrm{L})}\big) &= \big(\min(h_{\overline{o}}^{(\mathrm{L})}, \overline{h}_{\overline{o}}^{(\mathrm{L})}), \max(w_{\overline{o}}^{(\mathrm{L})}, \overline{w}_{\overline{o}}^{(\mathrm{L})})\big), \\ \big(\mathring{h}_{\overline{o}}^{(\mathrm{R})}, \mathring{w}_{\overline{o}}^{(\mathrm{R})}\big) &= \big(\min(h_{\overline{o}}^{(\mathrm{R})}, \overline{h}_{\overline{o}}^{(\mathrm{R})}), \max(w_{\overline{o}}^{(\mathrm{R})}, \overline{w}_{\overline{o}}^{(\mathrm{R})})\big). \end{aligned} \tag{2}$$

Similar as before, $\mathring{\mathcal{Y}}_{\overline{o}}$ contains all pixel positions $(h, w) \in \mathcal{I}$ that are enclosed by the object localization represented as $\mathring{\boldsymbol{y}}_{\overline{o}}$. Bringing all together, we define the gIoU between two localizations of the same object with identifier $\overline{o}$ as

$$gIoU_{\overline{o}} = \frac{|\mathcal{Y}_{\overline{o}} \cap \overline{\mathcal{Y}}_{\overline{o}}|}{|\mathcal{Y}_{\overline{o}} \cup \overline{\mathcal{Y}}_{\overline{o}}|} - \frac{|\mathring{\mathcal{Y}}_{\overline{o}} \backslash (\mathcal{Y}_{\overline{o}} \cup \overline{\mathcal{Y}}_{\overline{o}})|}{|\mathring{\mathcal{Y}}_{\overline{o}}|} \tag{3}$$

and the respective gIoU loss as

$$J^{\text{gIoU}} = \frac{1}{N_{\overline{\mathcal{O}}}} \sum_{\overline{o} \in \overline{\mathcal{O}}} (1 - gIoU_{\overline{o}}) \tag{4}$$

Table 1. **Localization label quality for noisy and refined data.** Mean intersection-over-union ($mIoU$) in % between ground truth data $\mathcal{D}_{\text{val}}$ and noisy data $\widetilde{\mathcal{D}}_{\text{val}}^{(\epsilon)}$, and between ground truth data $\mathcal{D}_{\text{val}}$ and refined data $\hat{\mathcal{D}}_{\text{val}}^{(\epsilon)}$ for different noise strengths ϵ are reported. The "single-pass"-row refers to single-pass `LLRN`-refined labels (our best method), while the "none"-row refers to noisy labels.

	Refinement method	Noise strength ϵ: 0	0.1	0.2	0.3	0.4	0.5
$mIoU$	single-pass	86.95	86.60	**85.85**	**84.45**	**82.55**	**79.50**
	none	**100.0**	**91.70**	84.15	77.50	71.35	65.60

Note that the first term in (3) is indeed the intersection over union between $\mathcal{Y}_{\overline{o}}$ and $\overline{\mathcal{Y}}_{\overline{o}}$ and the second term in (3) is a regularizer which penalizes the distance between $\mathcal{Y}_{\overline{o}}$ and $\overline{\mathcal{Y}}_{\overline{o}}$. In particular, the latter yields non-zero values for the special case when $\mathcal{Y}_{\overline{o}}$ and $\overline{\mathcal{Y}}_{\overline{o}}$ do not intersect with each other, where the first term in (3) ends up to be 0 for all non-intersecting cases. Whether $\overline{\mathcal{Y}}_{\overline{o}}$ or $\widetilde{\mathcal{Y}}_{\overline{o}}$ is used, depends on the `LLRN` type and iteration k.

2. Benchmarking and Qualitative Examples

For easier referencing and to establish a benchmark for localization label errors and their refinement, we make explicit our best $mIoU$ results in Tab. 1 and our best mAP results in Tab. 2. Further, we provide additional qualitative examples for the single-pass `LLRN` in Fig. 1.

References

[1] Ross Girshick. Fast R-CNN. In *Proc. of ICCV*, pages 1440–1448, Las Condes, Chile, Dec. 2015. 1

[2] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *Proc. of CVPR*, pages 658–66, Long Beach, CA, USA, June 2019. 1

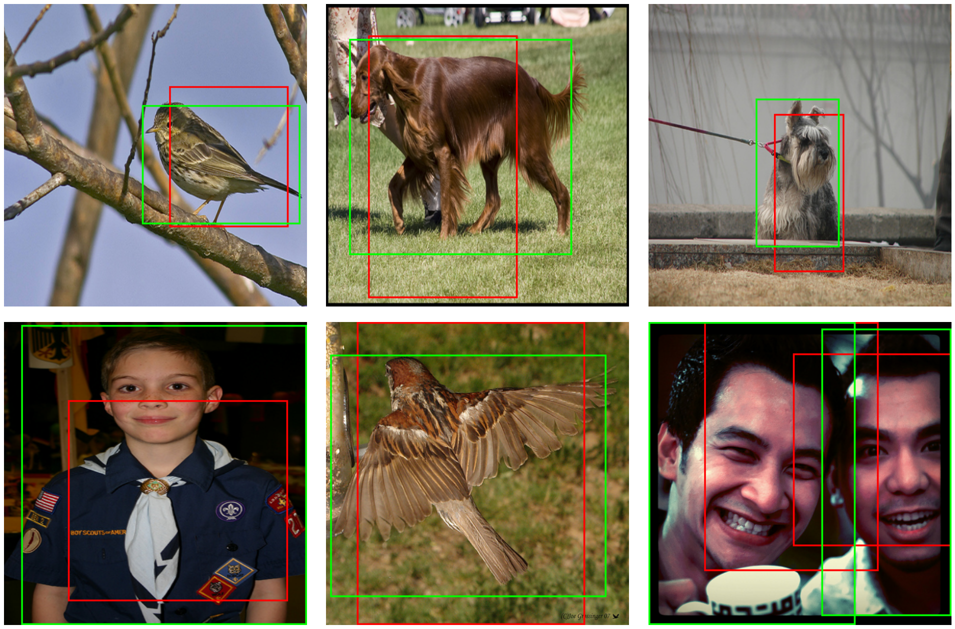

Figure 1. **More examples for our localization label refinement (single-pass LLRN approach)**. Our proposed framework takes a noisy localization label $\widetilde{\boldsymbol{y}}$ (red box) as input and outputs a refined localization label $\hat{\boldsymbol{y}}$ (green box).

Table 2. **Object detection performance on noisy and refined data**. Mean average precision (mAP_κ), $\kappa \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$, on $\mathcal{D}_{\text{val}}$ is reported in (%). We trained the `Cascade R-CNN` on noisy datasets $\widetilde{\mathcal{D}}_{\text{train}}^{(\epsilon)}$ and refined datasets $\hat{\mathcal{D}}_{\text{train}}^{(\epsilon)}$ with different (initial) noise strengths ϵ. The "single-pass"-rows refer to single-pass `LLRN`-refined labels (our best method), while the "none"-rows refer to noisy labels.

	Refinement method	Noise strength ϵ 0	0.1	0.2	0.3	0.4	0.5
$mAP_{0.5}$	single-pass	80.40	80.27	**80.77**	**80.03**	**80.03**	**79.10**
	none	**81.10**	**80.97**	80.53	79.33	77.63	75.23
$mAP_{0.6}$	single-pass	77.07	76.97	**77.17**	**76.53**	**76.37**	**75.10**
	none	**77.87**	**77.60**	76.67	74.57	71.53	66.90
$mAP_{0.7}$	single-pass	70.10	**70.17**	**70.20**	**69.23**	**68.33**	**66.37**
	none	**71.40**	70.13	67.73	63.03	55.90	45.33
$mAP_{0.8}$	single-pass	55.80	**55.80**	**55.40**	**54.10**	**51.57**	**47.50**
	none	**57.70**	54.30	46.97	35.83	23.17	13.30
$mAP_{0.9}$	single-pass	25.27	**25.00**	**23.90**	**21.77**	**18.23**	**12.13**
	none	**27.77**	19.43	7.50	2.63	1.03	0.40