Towards Robust Adaptive Object Detection under Noisy Annotations Supplementary Materials

Xinyu Liu[1] Wuyang Li[1] Qiushi Yang[1] Baopu Li[2] Yixuan Yuan[1,*]
[1]City University of Hong Kong [2]Baidu USA LLC
{xliu423-c, wuyangli2-c, qsyang2-c}@my.cityu.edu.hk, baopuli@baidu.com
yxyuan.ee@cityu.edu.hk

The Appendix contains following sections:

Appendix A is the proof of the first order approximation of gradient reconcilement.

Appendix B is the related works on domain adaptation under noisy source scenarios.

Appendix C is the full algorithm of the proposed NLTE.

Appendix D is additional experimental results and analysis of NLTE.

Appendix E is more qualitative results.

Appendix F is a discussion of the broader impact and limitations of the proposed task and method.

A. Proof of the First-order Approximation of Gradient Reconcilement

Theorem 1. For a domain adaptive object detector trained with noisy source annotations with a small learning rate α, the maximizing of the inner products of gradients provided by different samples

$$\arg \max_{\phi_f, \phi_{det}, \phi_{dis}} \left(G^s_{cln} \cdot G^s_{cpt} + G^s_{cln} \cdot G^t + G^s_{cpt} \cdot G^t\right), \quad (1)$$

can be approximated as the first-order meta-update between gradients over iterations:

$$(\phi_f, \phi_{det}) \leftarrow (\phi_f, \phi_{det}) + \lambda(\Delta\tilde{\phi}_f, \Delta\tilde{\phi}_{det}). \quad (2)$$

Proof. Without loss of generality, we use the situation of 2 iterations as the interval of meta updating. We first make the following definitions: Gradient at i respecting to the SGD updated model $\tilde{\phi}$ after i iterations:

$$\tilde{G}_i = \mathbb{E}_{x \in \mathbf{b}_i} \frac{\partial \mathcal{L}\left[(x^{s,t}, y^s); \tilde{\phi}\right]}{\partial \tilde{\phi}}. \quad (3)$$

Gradient respect to the model at the initial time ϕ:

$$G_i = \mathbb{E}_{x \in \mathbf{b}_i} \frac{\partial \mathcal{L}\left[(x^{s,t}, y^s); \phi\right]}{\partial \phi}, \quad (4)$$

here $\mathbf{b}_1$ represents the sampled mini-batch within the first step, which correspond to a linear combination of the mini-batches $\mathbf{b}^s_{cln}$, $\mathbf{b}^s_{cpt}$, and $\mathbf{b}^t$ sampled from $\mathcal{D}^s_{cln}$, $\mathcal{D}^s_{cpt}$, and $\mathcal{D}^t$, respectively.[1] Note that during training, same number of source and target images are fed into the model ϕ, the summation of source samples and target samples are equivalent, *i.e.*, $\mathbf{b}^s_{cln}$ + $\mathbf{b}^s_{cpt}$ = $\mathbf{b}^t$. However, as $\mathcal{D}^s_{cln}$ and $\mathcal{D}^s_{cpt}$ are implicitly drawn from different distributions $P^s_{X^s Y^s}$ and $P^s_{X^s \tilde{Y}^s}$, the ratio of $\mathbf{b}^s_{cln}$ and $\mathbf{b}^s_{cpt}$ are stochastic over iterations. Then, the Hessian at the initial point is:

$$H_i = \mathbb{E}_{x \in \mathbf{b}_i} \frac{\partial^2 \mathcal{L}\left[(x^{s,t}, y^s); \phi\right]}{\partial \phi^2}, \quad (5)$$

Following [5, 8], the second-order Taylor approximation of G_i is:

$$\begin{aligned} \tilde{G}_i &= \mathcal{L}'_i(\tilde{\phi}) \\ &= \mathcal{L}'_i(\phi) + \mathcal{L}''_i(\phi)\left(\tilde{\phi} - \phi_1\right) + O\left(\alpha^2\right) \\ &= G_i + H_i(\tilde{\phi} - \phi) + O(\alpha^2), \end{aligned} \quad (6)$$

by substituting $\tilde{\phi} - \phi = -\alpha \sum_{j=1}^{i-1} G_j$ and $\tilde{G}_j = G_j + O(\alpha)$ into the above equation, we have:

$$\tilde{G}_i = G_i - \alpha H_i \sum_{j=1}^{i-1} G_j + O\left(\alpha^2\right). \quad (7)$$

For the case where 2 iteration intervals:

$$\tilde{G}_2 = G_2 - \alpha H_2 G_1 + O\left(\alpha^2\right). \quad (8)$$

Regarding the gradient in Eq. (2) and substitute Eq. (8) in, we have:

$$\begin{aligned} \Delta\tilde{\phi} &= \tilde{\phi} - \phi = \alpha(\tilde{G}_1 + \tilde{G}_2) \\ &= \alpha(G_1 + G_2 - \alpha H_2 G_1 + O\left(\alpha^2\right)) \\ &= \alpha(G_1 + G_2) + \alpha^2(H_2 G_1) + O(\alpha^3). \end{aligned} \quad (9)$$

*Corresponding author.

[1]The mini-batches refer to the samples passed into the multi-class classifier, which correspond to the pooled RoI features.

Rewriting the second term:

$$\begin{aligned}\alpha^2(H_2G_1) &= \mathbb{E}_{1,2}\left[H_2G_1\right] \\ &= \mathbb{E}_{1,2}\left[H_1G_2\right] \\ &= \frac{1}{2}\left(\mathbb{E}_{1,2}\left[H_2G_1\right] + \mathbb{E}_{1,2}\left[H_1G_2\right]\right) \\ &= \frac{1}{2}\mathbb{E}_{1,2}\left[\frac{\partial}{\partial\phi}\left(G_1 \cdot G_2\right)\right].\end{aligned} \tag{10}$$

The equation can also be extended to k iterations by multiplying the equation by $\frac{k(k-1)}{2}$. Denoting the first iteration gradient respect to ϕ as G_1, the second iteration gradient respect to ϕ after 1 iteration as G_2, we have:

$$\begin{aligned}G_1 &= G^s_{cln,1} + G^s_{cpt,1} + G^t_1, \\ G_2 &= G^s_{cln,2} + G^s_{cpt,2} + G^t_2.\end{aligned} \tag{11}$$

Since only inner products of gradients from different types of samples instead of iterations are concerned, the iteration notations can be removed for clearer illustration, *i.e.*, $G^s_{cln,1}$ and $G^s_{cln,2}$ are both treated as G^s_{cln}. Thus, according to the distributive law of vector inner products, the inner product between G_1 and G_2 are:

$$\begin{aligned}G_1 \cdot G_2 &= (G^s_{cln} + G^s_{cpt} + G^t) \cdot (G^s_{cln} + G^s_{cpt} + G^t) \\ &= \|G^s_{cln}\|^2 + \|G^s_{cpt}\|^2 + \|G^t\|^2 \\ &\quad + \underbrace{G^s_{cln} \cdot G^s_{cpt} + G^s_{cln} \cdot G^t + G^s_{cpt} \cdot G^t}_{\text{The Inner Product Term}}\end{aligned} \tag{12}$$

Combining Eq. (9), Eq. (10), and Eq. (12), we show that using the residual of the model after multiple iterations as meta gradients is an approximation of the inner product of the gradients provided by same type of samples across iterations and different types of samples.

B. Related Works on Domain Adaptation with Noisy Annotated Source Dataset

Recently, some works attempted to achieve domain adaptation under noisy source annotations for the image classification task [3, 9, 14]. Shu *et al*. [9] designed a curriculum training strategy to select easy and transferable samples to jointly address label noise and feature noise. Han *et al*. [3] used offline curriculum learning as a more reliable strategy for selecting clean samples, meanwhile use proxy distribution search as a novel discrimination criterion. However, both methods assume the availability of noisy rate within the source dataset. Zuo *et al*. [14] designed a model-agnostic approach based on co-teaching [2], which utilized the outputs of two symmetrical domain adaptation models to explore noisy samples. However, the method is based on the small-loss strategy which may treat hard positives as noise. Meanwhile, it duplicates domain adaption models that may cause significantly increase in the storage demand and training time. Therefore, these related works could not be directly adopted to the noisy DAOD setting. To this end, we proposed NLTE for this challenging yet undeveloped task, which boosts the robustness of domain adaptive object detectors under noisy annotation scenarios effectively.

Algorithm 1 The training procedure of NLTE

Input: Noisy source dataset $\mathcal{D}^s = \{x^s, \tilde{y}^s\}$; Target dataset $\mathcal{D}^t = \{x^t\}$; Gradient reconcilement period κ; Untrained domain adaptive object detector $\phi = (\phi_f, \phi_{det}, \phi_{dis})$
Output: Trained noise-robust domain adaptive object detector ϕ
for $iter = 1$ **to** $maxiter$ **do**
  Feed x^s, x^t into ϕ, generate proposals $\mathbf{P}^s$, $\mathbf{P}^t$
  Potential Instance Mining:
  Select eligible proposals $\overline{\mathbf{P}}^s$, $\overline{\mathbf{P}}^t$ with Eq. (1)
  Morphable Graph Relation Module:
  Do intra-domain graph aggregation with Eq. (2)
  Generate global relation matrix with Eq. (3)-(4)
  Generate local relation matrix with Eq. (5)
  Compute $\mathcal{L}_{mgrm}$
  Entropy-Aware Gradient Reconcilement:
  Feed concatenated proposal features and logits into entropy-aware discriminator
  Compute $\mathcal{L}_{det}$, $\mathcal{L}_{dis}$, $\mathcal{L}^{EAGR}_{dis}$
  if iter % κ = 0 **then**
    Meta update (ϕ_f, ϕ_{det}) with Eq. (11)
  end if
end for
Return Well-trained cross-domain object detector ϕ

C. Full Algorithm

Algorithm 1 shows the training procedure of NLTE. The numbers of equations within the algorithm refer to those in the main paper.

D. Experiments

D.1. Details of Datasets with Synthetic Noise

For the Noisy Pascal VOC dataset, we manually generate synthetic noise that can be categorized into two groups: *miss-annotated* and *class-corrupted*. To simulate the manual labeling mistakes, for a dataset with $C+1$ classes (class 0 refers to background), we randomly select $r\%$ instances and substitute their original class label C_i to another label C_j, where $r\%$ is the noisy rate and $i \neq j$. If $j = 0$, then the instance is directly removed.

D.2. Adopting Other Noise-robust Learning Approaches

Besides SCE [11], CP [6], and GCE [12], we also tested adopting other noise-robust learning approaches into the

Table 1. Results (%) on Pascal VOC & Noisy Pascal VOC → Clipart1k under different noisy rates (NR).

Pascal VOC & Noisy Pascal VOC → Clipart1k																							
NR	Methods	aero	bcycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	hrs	bike	prsn	plnt	sheep	sofa	train	tv	mAP	Imprv.
0%	DAF	29.0	45.1	33.3	25.8	28.6	48.0	39.8	12.3	35.3	50.3	22.9	17.4	33.4	33.8	59.2	44.8	20.7	26.0	45.3	49.6	35.0	0.0
	+SR	18.2	30.8	28.9	25.1	26.6	42.7	33.3	2.8	30.5	27.4	17.9	11.0	25.6	32.2	48.0	36.4	16.3	18.8	41.0	43.9	27.9	-7.1
	+NLTE	39.1	50.3	33.6	34.7	35.0	40.5	44.2	5.9	36.8	45.8	23.1	17.3	31.8	39.5	60.7	45.4	17.9	28.4	49.0	51.3	**36.5**	**+1.5**
20%	DAF	34.0	39.1	32.0	27.3	32.2	39.3	38.9	2.9	34.9	44.9	20.6	14.2	30.8	36.6	53.8	43.8	17.6	23.6	42.8	46.1	32.8	0.0
	+SR	18.2	30.8	28.9	25.1	26.6	42.7	33.3	2.8	30.5	27.4	17.9	11.0	25.6	32.2	48.0	36.4	16.3	18.8	41.0	43.9	27.9	-4.9
	+NLTE	33.1	47.5	35.5	28.2	33.7	53.8	43.8	4.2	34.2	48.4	19.3	14.6	29.7	47.2	57.1	42.5	17.7	27.7	40.0	44.5	**35.1**	**+2.3**
40%	DAF	24.5	39.4	29.1	26.9	32.8	46.5	40.0	4.7	36.1	42.0	21.3	10.6	27.8	37.3	52.8	39.7	17.5	26.9	36.0	46.2	31.9	0.0
	+SR	15.3	43.3	30.7	15.3	29.8	42.4	32.3	5.0	26.4	41.7	17.2	18.5	26.5	37.8	50.6	42.0	18.0	16.9	38.1	38.5	29.3	-2.6
	+NLTE	32.8	45.5	30.8	29.8	35.7	43.2	43.0	6.4	32.7	45.9	19.8	10.8	31.1	43.4	56.4	43.3	19.6	24.8	42.5	43.9	**34.1**	**+2.2**
60%	DAF	29.4	33.5	29.7	29.0	27.7	39.5	38.0	2.7	31.9	41.5	19.8	12.9	30.2	37.0	49.7	37.2	12.8	25.5	40.8	44.2	30.6	0.0
	+SR	18.7	34.3	29.2	26.8	28.1	37.2	29.6	10.0	22.7	36.4	15.5	9.7	16.7	43.3	52.1	35.2	16.2	18.4	30.2	39.3	27.5	-3.1
	+NLTE	33.0	51.9	32.2	31.7	29.9	39.7	43.6	11.0	36.4	40.7	27.0	11.8	30.3	35.3	55.9	42.2	20.8	30.1	34.5	41.2	**34.0**	**+3.4**
80%	DAF	28.2	34.0	29.6	20.8	27.7	45.0	34.4	1.4	31.5	34.1	19.9	9.3	26.2	33.3	46.0	37.4	17.5	20.4	30.6	41.9	28.5	0.0
	+SR	20.9	36.9	19.8	21.7	26.6	38.8	26.2	4.7	26.5	28.9	16.2	7.3	19.3	49.4	46.0	38.2	15.3	8.6	36.9	31.9	26.0	-2.5
	+NLTE	36.0	45.4	33.5	30.3	27.3	40.5	40.6	2.6	28.3	51.7	20.4	9.5	30.8	43.1	56.6	42.1	17.7	23.3	31.2	38.4	**32.5**	**+4.0**

Table 2. Results on Noisy Pascal VOC → Watercolor2k under different noisy rates (NR).

Noisy Pascal VOC → Watercolor2k									
NR	Methods	bcycle	bird	car	cat	dog	prsn	mAP	Imprv.
20%	DAF	69.1	36.5	25.8	31.0	16.1	44.9	37.2	0.0
	+SCE	62.4	42.6	33.2	32.2	18.5	46.5	39.2	+2.0
	+NLTE	73.7	37.1	35.3	28.1	21.2	44.5	**40.0**	**+2.8**
	+NLTE(SCE)	78.4	39.2	38.5	28.2	27.9	49.3	**43.6**	**+6.4**
40%	DAF	68.0	32.9	20.5	19.8	13.6	39.4	32.4	0.0
	+SCE	64.5	36.6	37.8	14.1	14.0	42.8	35.0	+2.6
	+NLTE	75.7	37.2	32.5	22.6	24.3	43.1	**39.2**	**+6.8**
	+NLTE(SCE)	55.8	44.3	29.8	29.6	28.3	54.8	**40.5**	**+8.1**

DAF [1] framework. APL [4] combined active and passive loss functions to achieve a balance between under-fitting and over-fitting. However, replacing the cross-entropy loss [1] with different types of APL (NCE+RCE, NCE+MAE, NFL+RCE, NFL+MAE) will cause non-convergence *i.e.*, the mAP remains $5\% - 10\%$. A recent work SR [13] stated that any loss can be made robust to noisy labels by restricting the network output to the set of permutations over a fixed vector. However, as shown in Table 1, adopting SR in DAF [1] leads to performance drop on all settings. We conjecture that the loss of the detection framework is a summation of multiple losses, thus restricting the classification branch with sparse regularization individually may not be an effective solution.

D.3. Incorporation with Noise-robust Learning Approaches

Since the proposed NLTE is not aimed at correcting the noisy annotations but to explore their latent transferability for promoting the domain adaptation performance, we can incorporate it with previous noise-robust learning approaches, including SCE [11], CP [6], GCE [12], etc. However, most of them show negligible improvement and only SCE [11] benefits NLTE on the Noisy Pascal VOC → Watercolor2k setting. As shown in Table 2, we test the performance of implementing NLTE and SCE in conjunction on DAF and find it further improves the performance of the detector, which may attributing to the bias within the source domain is rectified to some extent and the difficulty of aligning noisy samples with NLTE is alleviated.

Table 3. Analysis of the gradient reconcilement period κ.

κ	1	2	4	6
mAP	34.8	35.1	34.6	34.1

Table 4. Analysis of the meta update weight λ.

λ	0.01	0.002	0.001	0.0005
mAP	34.3	34.1	35.1	34.5

D.4. Sensitivity Analysis of the Gradient Reconcilement Period

As illustrated in the main paper and Appendix A, The gradient reconcilement requires multiple iterations κ. To study the impact of κ, we conduct experiments on Noisy Pascal VOC → Clipart1k with noisy rate 20% and display the results in Table 3. Specially, $\kappa = 1$ is the original SGD training without reconcilement. We show that compared with the model without reoncilement, $\kappa = 2$ shows higher mAP 35.1, which is because gradients of distinct samples are encouraged to achieve coherence with gradient reconcilement. Therefore, both clean and noisy samples would be contributive towards learning a domain-invariant object detector. Besides, the performance drops as κ increases. The possible reason is that the accumulated gradients provided by noisy samples have dominated the training process, and harmonizing their directions could affect the clean samples. Therefore, we set $\kappa = 2$ in all experiments.

D.5. Sensitivity Analysis of the Meta Update Weight

To approximate the gradient reconcilement process, we utilize the first-order meta update for the model. To study the sensitivity of the weight λ of the weighted combination of the model before and after κ iterations, different values of λ are tested on the Noisy Pascal VOC → Clipart1k with

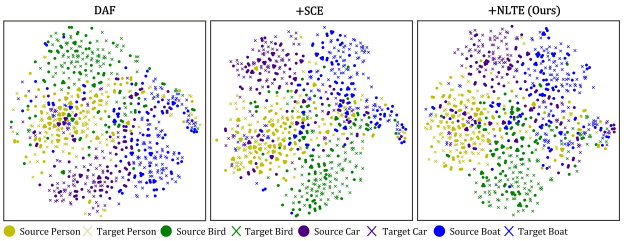


Figure 1. Visualization of features obtained by different models in the Noisy Pascal VOC → Clipart1k under 20% noisy rate. The dots and cross marks refer to source and target samples, and each color refers to a different class.

noisy rate 20%. As illustrated in Table 4, the model is quite robust to the hyperparameter, and we set it to 0.001 in all our experiments.

D.6. Visualization of the Feature Distributions

The t-SNE [10] visualizations of the feature distributions are presented in Fig. 1. The features are extracted from proposals that correspond different instances and we select four classes for clearer interpretation. Although DAF [1] and SCE [11] could align source and target domains, they suffer from miss classification caused by noisy annotations as their are features are more mixed. In contrast, NLTE shows better ability in distinguishing different classes meanwhile preserves the ability of adaptation as different domain samples are well aligned.

E. Extended Qualitative Results

We provide extended qualitative results on Noisy Pascal VOC → Clipart1k with 20% noisy rate in Fig. 2, Noisy Pascal VOC → Watercolor2k with 20% noisy rate in Fig. 3. Compared with other noise-robust learning approaches, DAF+NLTE can categorize objects in the target domain more accurately, indicating its robustness towards noisy source annotations. We also provide qualitative detection results on Cityscapes → Foggy Cityscapes in Fig. 4, and display the source only results and the ground truth for clearer illustration. It is observed that DAF+NLTE consistently outperforms the source only model. It can not only alleviate false negative instances but also predict the corresponding class labels correctly. This scenario demonstrates the effectiveness of NLTE in addressing the noisy in natural scenarios and benefiting domain adaptive object detection. However, there remain failure detections, which may due to the large semantic gap between the original label and the corrupted label, or due to the incapacity of the feature alignment between source and target domains.

F. Broader Impact and Limitations

The proposed method detects objects across domains even under noisy source annotations, which may reduce the burden of labeling and one may use abundant coarse web-crawled images to train a domain adaptive object detector. Meanwhile, our method could also be used for estimating the bias level of the annotated dataset. However, our work is restricted to the setting where source and target domain share the same label space. Thus our future work will also include the extension to open-set scenarios.

References

[1] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 3, 4, 5, 6

[2] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv preprint arXiv:1804.06872*, 2018. 2

[3] Zhongyi Han, Xian-Jin Gui, Chaoran Cui, and Yilong Yin. Towards accurate and robust domain adaptation under noisy environments. *arXiv preprint arXiv:2004.12529*, 2020. 2

[4] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *ICML*, pages 6543–6553, 2020. 3

[5] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 1

[6] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 2, 3, 5, 6

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 7

[8] Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021. 1

[9] Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Transferable curriculum for weakly-supervised domain adaptation. In *AAAI*, volume 33, pages 4951–4958, 2019. 2

[10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(11), 2008. 4

[11] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019. 2, 3, 4, 5, 6

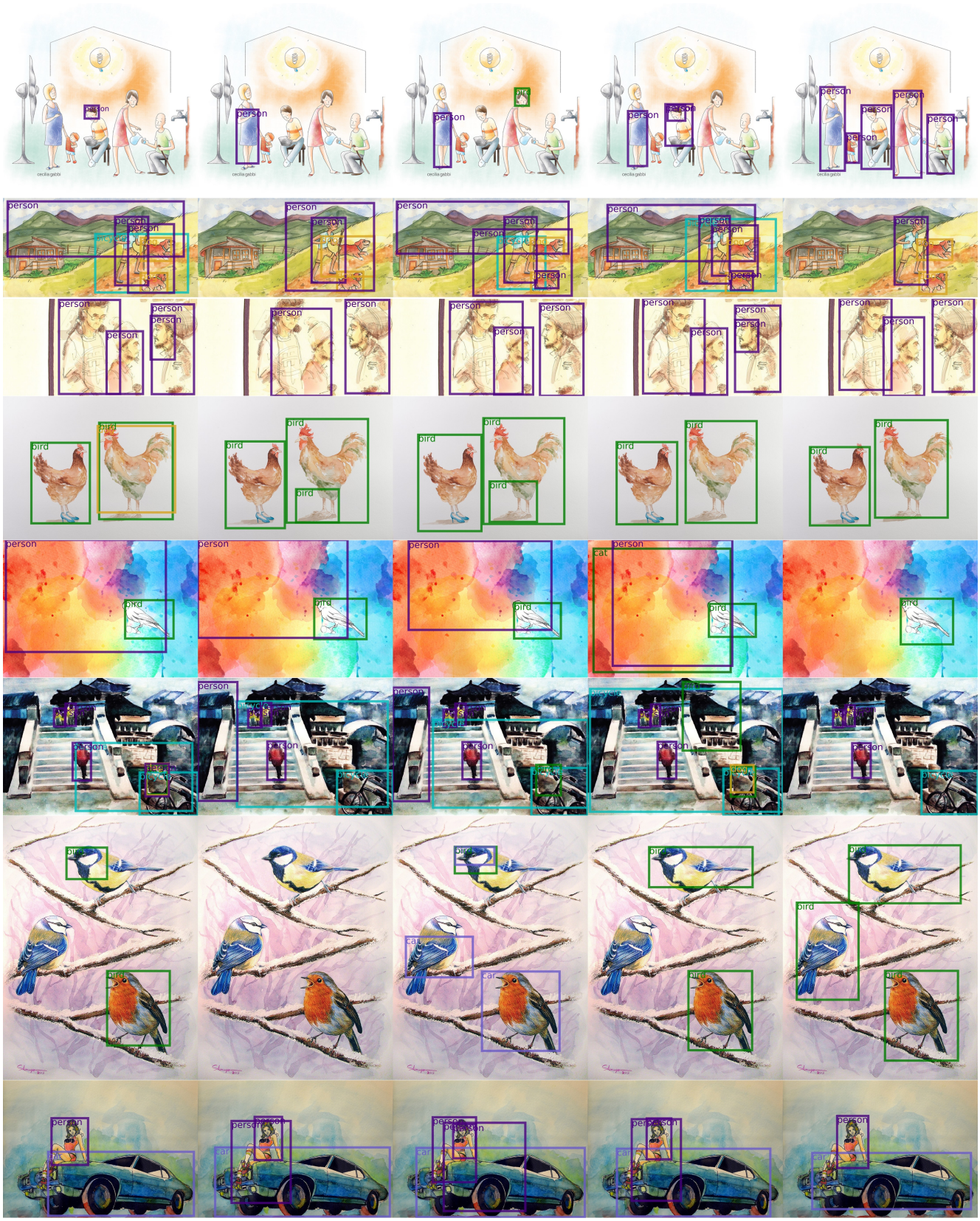
[12] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018. 2, 3, 5, 6

[13] Xiong Zhou, Xianming Liu, Chenyang Wang, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Learning with noisy labels via sparse regularization. In *ICCV*, pages 72–81, 2021. 3

[14] Yukun Zuo, Hantao Yao, Liansheng Zhuang, and Changsheng Xu. Seek common ground while reserving differences: A model-agnostic module for noisy domain adaptation. *IEEE Trans. Multimedia*, 2021. 2

(a) DAF [1] (b) +SCE [11] (c) +CP [6] (d) +GCE [12] (e) +NLTE (Ours)

Figure 2. Qualitative results with noisy rate 20% on Clipart1k.



(a) DAF [1] (b) +SCE [11] (c) +CP [6] (d) +GCE [12] (e) +NLTE (Ours)

Figure 3. Qualitative results with noisy rate 20% on Watercolor2k.

(a) Source Only [7] (b) DAF+NLTE (Ours) (c) Ground Truth

Figure 4. Qualitative results on Foggy Cityscapes.