ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414355

NLKD: USING COARSE ANNOTATIONS FOR SEMANTIC SEGMENTATION BASED ON KNOWLEDGE DISTILLATION

Dong Liang *Yun Du* *Han Sun* *Liyan Zhang* *Ningzhong Liu* *Mingqiang Wei*

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence,
Collaborative Innovation Center of Novel Software Technology and Industrialization

ABSTRACT

Modern supervised learning relies on a large amount of training data, yet there are many noisy annotations in real datasets. For semantic segmentation tasks, pixel-level annotation noise is typically located at the edge of an object, while pixels within objects are fine-annotated. We argue the coarse annotations can provide instructive supervised information to guide model training rather than be discarded. This paper proposes a noise learning framework based on knowledge distillation NLKD, to improve segmentation performance on unclean data. It utilizes a teacher network to guide the student network that constitutes the knowledge distillation process. The teacher and student generate the pseudo-labels and jointly evaluate the quality of annotations to generate weights for each sample. Experiments demonstrate the effectiveness of NLKD, and we observe better performance with boundary-aware teacher networks and evaluation metrics. Furthermore, the proposed approach is model-independent and easy to implement, appropriate for integration with other tasks and models.

Index Terms— Noisy label, knowledge distillation, semantic segmentation

1. INTRODUCTION

Fig. 1. Fine and coarse annotations on Cityscapes [1]. Left and right columns show fine annotations and corresponding coarse annotations, respectively.

This work is partially supported by AI+ Project of NUAA (XZA20003), National Science Foundation of China (61772268). Dong Liang and Yun Du contribute equally to this work. Han Sun is the corresponding author.
{liangdong, muyun, sunhan}@nuaa.edu.cn

Although deep neural networks have already achieved tremendous success in semantic segmentation, their performance suffers from noisy labels in training data. As illustrated in Fig.1, the annotation in the segmentation task can be roughly classified into the fine one and the coarse one. Since semantic segmentation annotation requires assigning a class label to each pixel of the image, pixel-level annotation of object edges is easily misaligned. Moreover, high-quality annotations are costly and time-consuming. As reported by the Cityscapes dataset [1], an image with a shape of 2048×1024 costs 1.5 hours to get a fine label while only 7 minutes to generate a coarse one. As a result, most of the deep learning models in the industry have to learn from a lot of noisy data.

Recent studies demonstrate three ways to cope with noisy data. 1) pre-training on coarsely-labeled images and then fine-tuning on finely-labeled ones; 2) training robust models on noisy data; 3) detecting and discarding noise data to get a clean dataset.

[2] analyzes the effect of data quality on semantic segmentation and proposes the idea of pre-training with noisy data and fine-tuning with finely annotated data. [3] designed O2U-Net with intuition, suggested that training with different learning rates makes the model transfer from overfitting to underfitting cyclically, and the average sample losses can be the indicator of the probability of label noise. [4] proposed a joint optimization framework of learning DNN parameters and estimating true labels. And it can correct labels during training by the alternating update of network parameters and labels. [5] adopt gradient descent on sound data and learning-rate-reduced gradient ascent on bad data to avoid memorizing noisy data.

Knowledge distillation [6] has been widely used in model compression and transfer learning. This framework introduces soft targets generated by the teacher network and the student network and considers it an additional optimization objective, enabling knowledge transfer from the teacher to the student. [7] introduced knowledge distillation methods into the task of learning with noise, including a small clean dataset and label relations in a knowledge graph. [8] train student networks using pseudo-labels generated by teacher networks and

cyclically swap teacher and student networks' roles.

Inspired by [2, 5, 9], we attempted to train a robust model using noisy annotation data. To achieve this goal, we try to suppress noise in the coarse annotations by a re-weighting strategy based on knowledge distillation to transfer the teacher model's knowledge to the student model, which allows the student to achieve better performance. Specifically, the models (both student and teacher networks) generate pseudo-labels for each sample and re-weight the loss by mIoU and boundary-aware score calculated between the pseudo-labels and the accurate annotations. The final weights are composed of student's and teacher's weight, the proportion of which will be adjusted over the training process. In the first few epochs, the weights calculated by the student network will be fluctuating and unreliable because of its unfitting. However, as training epochs increase, the proportion of student's weights increases progressively, which is more conducive to performance.

2. METHODOLOGY

2.1. Reweight is better than discarding

Many methods are appropriate for image classification but not effective for semantic segmentation. As for the classification tasks, the gradient returned by the noisy annotations is completely wrong, which is negative for the model optimization process. However, for semantic segmentation, the wrong pixel annotations usually appear at the edge of objects. There are still many correctly annotated pixels inside the object, which is helpful to model optimization. As illustrated in Tab.1, the approach with detect-discard strategy is unfavorable for the semantic segmentation task, and we consider re-weight the sample loss to deal with this problem.

Table 1. The effect of coarse segmentation data. The performance of the model is reduced by supplementing additional coarse data. However, the performance is improved if the loss weights of the coarse annotations are reduced to half that of the fine sample.

Num images		Method	Metric	
clean	coarse	Loss Reweight	mIoU	BF
4000	0	✗	86.88	**59.77**
4000	500	✗	86.14	55.79
4000	500	✓	**87.44**	59.74

2.2. Generate the sample weights by evaluating the quality of pseudo-labels

Learning the object boundary is quite challenging for segmentation tasks, while most of the noisy pixels are located at the boundary. We prefer to calculate a sample-level loss weight to suppress the noise of the coarse annotations, making the model tend to learn from the fine-labeled cases. The weights are derived by evaluating the differences between ground-truth and pseudo-labels generated by the teacher network and the student network, so we need to consider various evaluation indicators. Since the noise at the boundaries tends to be very severe, we consider boundary F1-measure [10] as one part of the loss weights of the samples. The Boundary F1-measure (BF) could be formulated as follows:

$$P^c = \frac{1}{|B_{ps}|} \sum_{z \in B_{ps}^c} \left[d\left(z, B_{gt}^c\right) < \theta \right] \tag{1}$$

$$R^c = \frac{1}{|B_{gt}|} \sum_{z \in B_{gt}^c} \left[d\left(z, B_{ps}^c\right) < \theta \right] \tag{2}$$

$$BF_1^c = \frac{2 \cdot P^c \cdot R^c}{R^c + P^c} \tag{3}$$

where B_{gt}^c is the boundary map of the ground truth segmentation map for class c, similarly B_{ps}^c is the contour map for the predicted segmentation map $S_p s^c$. $[z]$ is the Iverson bracket notation. It converts any logical proposition into a one if the proposition is satisfied and 0 otherwise. **BF** indicates whether the predicted boundary point matches the ground-truth boundary within a distance error tolerance θ.

However, BF relies on a distance threshold. If there are no pixels within the distance tolerance range, it tends to be very low or even close to zero as the Equ.3. In extreme cases, using a single metric BF can make the weighting strategy tend to discard samples. As in our analysis in Sec.2.1, discarding samples is detrimental to semantic segmentation, so we also take mIoU into account.

mIoU indicates the intersection-over-union between the predicted map and ground-truth pixel, averaged over all classes. However, it only captures the relationship between the pixel sets and cannot describe the matching of the target boundaries. Yet BF is complementary with mIoU as it more carefully takes the contours into account. As a result, we considered both measures simultaneously to calculate the mean values of mIoU and BF, which we named **BFmIoU**.

The experiment results are shown in Tab.2, both the individual metric are effective in improving the model performance. However, the group using BFmIoU has a more considerable performance boost than the BF group and mIoU group. BFmIoU can focus on the boundary of annotations without discarding the sample, which is an appropriate metric for calculating weights.

2.3. Obtaining dynamic weights by knowledge distillation

In this section, we proposed the re-weight strategy based on knowledge distillation for learning with noisy annotations. Fig.2 shows the pipeline of this strategy.

Firstly, both the student network and teacher network generate prediction maps for each image, then the segmentation loss is calculated as usual. However, before the gradient back-propagation, the loss needs to be re-weighted for each sample with the proposed NLKD method. The weight is jointly determined between the annotations and prediction maps. After

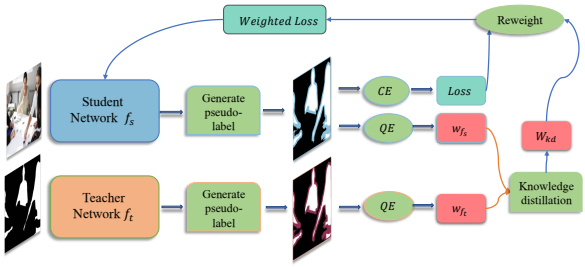


Fig. 2. The proposed distillation framework. Firstly, the student and teacher network generate pseudo-labels for each image. CE means cross-entropy loss and QE means quality evaluation. CE module calculates the $Loss$ and QE module generates the weights of each sample by evaluating the differences between annotations and pseudo-labels. The distilled weights W_{kd} re-weight the $Loss$ and then the $Weighted\ Loss$ updates the parameters of the student model.

calculating the teacher and student models' weights for each sample separately, NLKD generates the final weight in a distilled weighted manner. The student network updates parameters by optimizing the weighted loss.

There are a student network f_s and a teacher network f_t, the segmentation loss could be formulated as the pixel-wise cross-entropy loss:

$$L_s(x,y) = -\sum_{h=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C} y_{wh} log(f_s(x_{wh})) \tag{4}$$

Where H and W denote the height and the width of the input image, and C is the number of segmentation classes. The NLKD could be formulated as:

$$w_f(x,y) = (BF(f(x),y) + mIoU(f(x),y))/2 \tag{5}$$

$$W_{kd}(x,y) = \lambda w_{f_s}(x,y) + (1-\lambda) w_{f_t}(x,y) \tag{6}$$

$$L_b(X_b,Y_b) = \sum_{i=1}^{b} W_{kd}(x_i,y_i) L_s(x_i,y_i) \tag{7}$$

Where b is the batch size and λ is the parameter used to adjust student and teacher weights. Since the student network does not fit well at the beginning of training, we set a warm-up strategy to ensure the weights' stability and reliability. In practice, we set λ to 0 in the beginning epochs. After the warm-up period, the student network can generate good pseudo-labels and calculate sample weight, the value of λ increases.

For a batch of samples, we calculate the weighted segmentation loss of each batch with NLKD strategy (Equ.7), the student network updates parameters by minimizing weighted segmentation loss.

3. EXPERIMENTS

3.1. Implementation Details

In general, we select U-Net [11] and Deeplabv3+ [12] as student network, OCRNet [13] with HRNet [14] backbone and PointRend [15] as the teacher network. All experiments are performed on four GTX2080Ti GPUs, we set the optimizer to Ranger [16] with a weight decay $5e^{-4}$. The learning rate starts at 0.1 and changes to 0.5 times the original rate every 25 epochs. We set the number of training epochs to 100 and batch size to 32 for all trials. Besides, we initialize λ to 0 during the beginning 10 epochs, linearly increases λ to 0.5 through the remaining 90 epochs.

3.2. Evaluation on benchmarks

We selected 2 finely-labeled datasets and 2 coarsely-labeled ones to evaluate the effectiveness of our re-weight strategy.

Supervisely Fine. Supervisely [17] is a dataset for portrait segmentation, which is more finely labeled than other datasets such as COCO and VOC. We used 4500 images as a training set, 500 images as a validation set, and the remaining 711 images as a test set. The original version of Supervisely is named Supervisely Fine.

Supervisely Manual Coarse. We randomly perform dilation and erosion operations on the fine annotations to simulate the noisy boundary. To simulate the real noisy data and name this as Supervisely Manual Coarse, we did this for both the training set and the validation set

Cityscapes Fine. The Cityscapes dataset has clean annotations shown in Fig.1. We name this Cityscapes Fine. There are 2,975 images in Cityscapes Fine.

Cityscapes Official Coarse. The Cityscapes dataset has official coarse annotations. We name this Cityscapes Official Coarse. There are 2,975 images in Cityscapes Official Coarse.

We validate the effectiveness of NLKD strategy on both coarse-labeled and finely-labeled datasets and report the performance on the same test set. Comparison results on different benchmark are shown in Tab.3 with two metrics used for semantic segmentation: Mean IoU (mIoU) and Boundary F1-measure (BF).

We observe that the proposed NLKD strategy obtains consistent performance gains. By decreasing the weight of noisy samples, the performance on noisy datasets obtained improvements. In particular, on Cityscapes Official Coarse dataset, we improve mIoU by 2.77 and BF score by 8.96. Moreover, we also achieve an improvement on the finely-annotated dataset. On Supervisely Fine dataset, we got a 0.66 improvement on the mIoU metric.

3.3. Ablation Study

Tab.2 illustrates the performance improvement of our proposed method at different noise ratios. As for the Method column, **Reweight** means to re-weight the sample only using the teacher model since the teacher model will not update the parameters so that the sample weights are constant during the training process. **KD** means to combine the knowledge distillation framework. If combined, the teacher model and the student model jointly re-weight the sample, the sample weights

Table 2. Ablation experiment with different noisy ratio

	Method		Performance of student network (mIoU/BF)					
Metric	Reweight	KD	0%	10%	20%	30%	40%	50%
-			87.10/60.01	86.14/55.79	86.13/52.37	84.64/48.62	84.60/44.30	84.74/43.96
mIoU	✓		-	86.69/57.60	85.89/53.93	86.38/52.71	85.03/50.02	85.13/43.91
BF	✓		-	87.04/59.34	86.60/58.05	85.99/57.53	85.63/53.48	84.45/50.14
BF	✓	✓	-	87.28/59.51	86.85/**58.45**	86.20/57.39	85.86/54.63	85.63/54.12
BFmIoU	✓		-	86.62/57.31	86.82/57.31	**87.24**/58.46	87.08/57.89	84.95/46.63
BFmIoU	✓	✓	**87.76/61.12**	**87.44/59.74**	**86.94**/57.95	86.77/**58.55**	**87.41/59.67**	**86.80/58.03**

Table 3. Evaluate NLKD on public datasets

Dataset	Method	mIoU	BF
Supervisely Fine	baseline	87.10	60.01
Supervisely Fine	NLKD	**87.76(+0.66)**	**61.12(+1.11)**
Supervisely Manual Coarse	baseline	84.74	43.96
Supervisely Manual Coarse	NLKD	**86.80(+2.06)**	**58.03(+14.07)**
Cityscapes Fine	baseline	67.97	45.50
Cityscapes Fine	NLKD	**69.51(+1.54)**	**48.16(+2.66)**
Cityscapes Official Coarse	baseline	59.79	34.01
Cityscapes Official Coarse	NLKD	**62.56(+2.77)**	**42.98(+8.96)**

can change dynamically by fusing the two models' weights. The student network will keep updating the parameters during the training process.

We observe that the weights calculated under the KD framework perform better. This is because in the later stages of the training process, the loss weights generated by the teacher model have been fitted fairly well, and the student model's own knowledge becomes effective, similar to a self-learning approach of downweighting noisy samples. And different metrics also bring different performance gains. The results show that BF scores and BFmIoU improve more significantly than the mIoU group because BF and BFmIoU take boundary pixels into account.

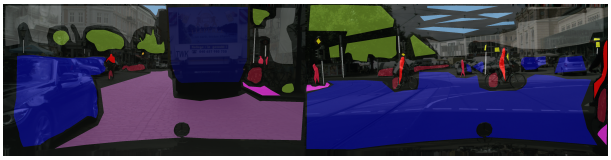
Table 4. Comparison on different networks

Student	Teacher	Method	mIoU	BF
U-Net [11]	-	baseline	84.74	43.96
U-Net [11]	OCRNet [13]	NLKD	85.16(+0.42)	48.92(+4.96)
U-Net [11]	PointRend [15]	NLKD	**86.80(+2.06)**	**58.03(+14.07)**
DeepLabv3+ [12]	-	baseline	86.70	50.41
DeepLabv3+ [12]	OCRNet [13]	NLKD	87.59(+0.89)	56.12(+5.71)
DeepLabv3+ [12]	PointRend [15]	NLKD	**87.95(+1.25)**	**58.06(+7.65)**

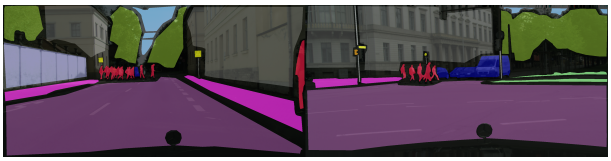
Table 5. Comparison on different method

ModelA	ModelB	Method	mIoU(ModelA)	BF(ModelA)
U-Net	-	baseline	84.74	43.96
U-Net	DeepLabv3+	Decoupling [18]	84.12	43.41
U-Net	DeepLabv3+	Co-teaching [19]	84.14	44.57
U-Net	DeepLabv3+	Co-teaching+ [20]	83.16	41.56
U-Net	DeepLabv3+	NLKD	**85.92**	**51.63**

Tab.4 explores the effectiveness of the strategy in different student networks and teacher networks. The results show that in the case where the same student model is all U-Net, both the teacher models can improve performance. However, PointRend improves more significantly. We believe that PointRend is better for object boundaries, providing more knowledge of the pixel weights of the boundary points. Tab.5 illustrates the comparison results with different approaches, [18, 19, 20] discard samples so their performance is inferior to the NLKD strategy. Fig.3 shows the cases with different weight.



(a) The cases with lowest weight



(b) The cases with highest weight

Fig. 3. Annotation with different weights on Cityscapes Official Coarse. The first row is coarsest samples with lowest sample weight and the second row is with highest weight.

4. CONCLUSION

In this paper, we propose a noise-robust learning framework based on knowledge distillation and named NLKD, which reduces the performance damage from noisy annotations in a simple and effective way. NLKD generates the pseudo-labels and then calculates the loss weights for each sample by jointly evaluating the quality of annotations with the teacher and student models. With the weighted loss, the student model tends to learn the finely labeled samples. The results demonstrate the effectiveness of this framework and the significance of boundaries in the semantic segmentation task. Comparison results on different benchmarks show that the networks are replaceable in the framework, fully illustrating the flexibility of this framework.

5. REFERENCES

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[2] Aleksandar Zlateski, Ronnachai Jaroensri, Prafull Sharma, and Frédo Durand, "On the importance of label quality for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1479–1487.

[3] Jinchi Huang, Lie Qu, Rongfei Jia, and Binqiang Zhao, "O2u-net: A simple noisy label detection approach for deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3326–3334.

[4] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.

[5] Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor W Tsang, and Masashi Sugiyama, "Sigua: Forgetting may make learning with noisy labels more robust," ICML, 2020.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *stat*, vol. 1050, pp. 9, 2015.

[7] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1910–1918.

[8] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10687–10698.

[9] Jingchao Liu, Ye Du, Qingjie Liu, and Yunhong Wang, "Variance loss: A confidence-based reweighting strategy for coarse semantic segmentation," *arXiv preprint arXiv:2009.05205*, 2020.

[10] Gabriela Csurka, Diane Larlus, Florent Perronnin, and France Meylan, "What is a good evaluation measure for semantic segmentation?.," in *BMVC*, 2013, vol. 27, p. 2013.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[13] Yuhui Yuan, Xilin Chen, and Jingdong Wang, "Object-contextual representations for semantic segmentation," in *ECCV*, 2020.

[14] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2019.

[15] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick, "Pointrend: Image segmentation as rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9799–9808.

[16] Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang, "Gradient centralization: A new optimization technique for deep neural networks," in *European Conference on Computer Vision*. Springer, 2020, pp. 635–652.

[17] "Supervisely person," https://supervise.ly/explore/projects/supervisely-person-dataset-23304/datasets.

[18] Eran Malach and Shai Shalev-Shwartz, "Decoupling "when to update" from "how to update"," in *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 960–970.

[19] Masashi Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *NeurIPS*, 2018.

[20] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama, "How does disagreement help generalization against label corruption?," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7164–7173.