SAE: ESTIMATION FOR TRANSITION MATRIX IN ANNOTATION ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The transition matrix plays a critical role in label-noise learning tasks, which refers to the transition from clean labels to noisy labels. The majority of recent methods for inferring the transition matrix concentrate on the manually hand-crafted label noise but with bearing the high cost of time and labor. In light of this, several straightforward and effective algorithms are introduced for automatically annotating the label noise. However, the automatic annotation algorithms easily generate wrong pseudo labels for similar semantic categories. Moreover, a special instance-dependent transition matrix is launched due to the mapping from a specific category to other similar categories during the annotation process. To address this issue, we propose a semantic adaption estimator (SAE) to indirectly infer the instance-dependent transition matrix. Specifically, we decouple the original instance-dependent transition matrix to several easy-to-estimate semantic-dependent transition matrices by introducing a semantic adaption loss function. In this way, the original datasets can be decoupled into some simple semantic regions. Then the instance-dependent transition matrix can be built from multiple learned semantic-dependent matrices. Empirical evaluations on two real-world datasets (*i.e.*, S3DIS and ScanNet) demonstrate the superior performance of our method, in comparison with the state-of-the-art.

1 INTRODUCTION

Deep learning techniques have shed light on the point cloud semantic segmentation tasks. The majority of existing models are derived based on large annotated training datasets Qi et al. (2017a;b); Li et al. (2018); Huang et al. (2018); Wang et al. (2018); Ren et al. (2022); Wang et al. (2019b). Nevertheless, accurately annotating large datasets is generally expensive and sometimes impracticable, especially for the point cloud data which is unordered, unstructured, and non-uniform Hu et al. (2021). To address this issue, the researchers have turned their appetites to the cheap datasets with label noise Wu et al. (2020); Veit et al. (2017); Vahdat (2017); Han et al. (2018a); Guo et al. (2018); Tanaka et al. (2018); Ma et al. (2018); Jiang et al. (2018); Ren et al. (2018); Han et al. (2018b); Yu et al. (2019); Liu & Guo (2020); Xu et al. (2019); Yu et al. (2018); Thekumparampil et al. (2018); Patrini et al. (2017); Goldberger & Ben-Reuven (2016); Kremer et al. (2018); Reed et al. (2014). Up-to-date results demonstrate that the label noise has the tendency in notably degenerating the performance of semantic segmentation models, as the deep models are capable of over-fitting the label noise Wei et al. (2020).

Existing solutions with noisy labels generally fall into two categories: model-free and model-based algorithms. The former one refers to employing several heuristic ways to decrease the side reactions of label noise, such as the robust regularization Guo et al. (2018); Veit et al. (2017); Vahdat (2017); Li et al. (2020; 2017), the samples selection scheme Yu et al. (2019); Yao et al. (2020a); Ren et al. (2018); Malach & Shalev-Shwartz (2017); Jiang et al. (2018); Han et al. (2020; 2018b), and the loss designation Ren et al. (2018); Ma et al. (2020); Lyu & Tsang (2019); Wang et al. (2019a); Amid et al. (2019b;a); Zhang & Sabuncu (2018); Ghosh et al. (2017). Their credibility is usually uncertain due to the failure to model the label noise, although several tremendous progress has been made. Alternatively, the model-based algorithms aim to converge to the optimal solution defined on the clean samples, where the model is learned by exploring noisy data systematically.

The transition matrix $T(x) \in [0,1]^{c\times c}$ plays an essential role in model-based algorithms that transfers the model from clean domain to noisy domain, where $T_{ij}(x) = P(\widetilde{\mathcal{Y}} = j \mid \mathcal{Y} = i, \mathcal{X} = x)$. We further set $P(A)$ as the probability of the event A, $\mathcal{X}$ as the data feature space, $\mathcal{Y}$ and $\widetilde{\mathcal{Y}}$ as the clean and noisy label space, respectively. Then we assume the x is a data example of $\mathcal{X}$, T_{ij} is the probability of flipping the clean label $\mathcal{Y} = i$ to the noisy label $\widetilde{\mathcal{Y}} = j$. Given $T(x)$, the basic idea is that the clean label $\mathcal{Y}$ can be inferred by utilizing the transition matrix $T(x)$ and noisy label posterior (which can be estimated from noisy label $\widetilde{\mathcal{Y}}$). Yet the transition matrix $T(x)$ is unspecified which incurs the difficulties in learning $T(x)$ from noisy data. Despite several valid techniques have been explored such as the robust loss designation, the regularization scheme, and meta learning, semi-supervised learning schemes, it is still difficult to model the instance-dependent label noise transition matrix, especially for the noise labels caused by annotation algorithms.

Motivated by the fact that the instances are annotated in accordance with their semantic features rather manually hand-crafted, we in this paper propose a semantic adaption estimator (SAE) to indirectly infer the instance-dependent transition matrix to address the aforementioned issue. Specifically, we decouple the original instance-dependent transition matrix $T(x)$ to several easy-to-estimate semantic-dependent transition matrices $T(x)_i$ $(i = 1, 2, \ldots, N)$ by introducing a semantic adaption loss function. We assume that the probability of flipping the semantic-dependent transition matrix $T(x)_i$ to another semantic-dependent transition matrix $T(x)_j$ is independent, and the contribution of semantic-dependent transition matrices to $T(x)$ is identical [1]. In this case, the transition matrix $T(x)$ can be approximately reconstructed from several simple semantic-dependent transition matrices $T(x)_1, T(x)_2, \ldots, T(x)_N$ that are learned by the semantic adaption estimators. The whole process can be found in Figure. 1. Moreover, the empirical evaluations on two real-world datasets (i.e., S3DIS and ScanNet) demonstrate the superior performance of our method, in comparison with the state-of-the-art.

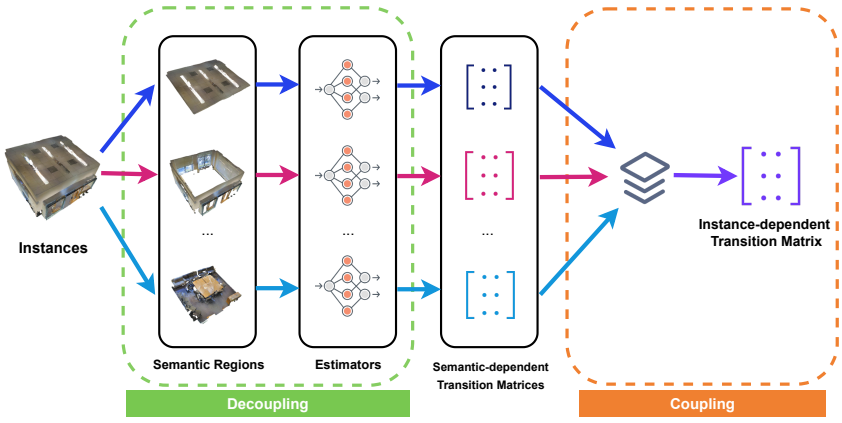

Figure 1: Framework of our semantic adaptation estimator (SAE). It consists of two key components: (1) The interpretable compositional layers work with semantic adaptation loss, which are responsible for decoupling the input scene datasets to different semantic regions. The estimators take these semantic regions as input and model semantic-dependent transition matrices. (2) The instance-dependent transition matrix for each instance can be approximated by a coupled of the semantic-dependent transition matrices.

The rest of the paper is organized as follows. In Section 2, we discuss how to learn semantic-dependent transition matrices. In Section 3, we describe the details of SAE and prove the statistically consistent of the SAE. In Section 4, we provide empirical evaluations of the SAE. In Section 5, we conclude our paper.

[1]The assumption holds since the similar semantic-dependent categories are easy to be annotated, and noisy labels from semantic-independent categories are few.

2 ESTIMATING TRANSITION MATRIX IN THE LABEL-NOISE LEARNING

In this section, we give a thorough description of the transition matrix with a mathematical formulation. First, we show the problem setup of the objective function in the label noise point cloud. To learn a robust model from noisy labels, we then give the definition of the transition matrix. Finally, the mathematical form of the transition matrix estimation is given.

2.1 PROBLEM SETUP

In this paper, we set the point cloud sample as X with the corresponded semantic label being Y, and we have $(X, Y) \in \mathcal{X} \times \{1, \ldots, C\}$, where $\mathcal{X}$ represents the data feature space, and C is the size of semantic categories. Let D be the clean training data having $\mathcal{D} = (\mathcal{X}, \mathcal{Y})$, $\widetilde{\mathcal{D}}$ be the distribution of the noisy pair $(\mathcal{X}, \widetilde{\mathcal{Y}})$ with $\widetilde{\mathcal{Y}}$ being the noisy semantic label space. We further detail $\widetilde{\mathcal{D}}$ as $\widetilde{\mathcal{D}} = \{(x_i, \widetilde{y}_i)\}_{i=1}^{n}$, where $(x_i, \widetilde{y}_i)$ is a subset of noisy pair $(\mathcal{X}, \widetilde{\mathcal{Y}})$ Yao et al. (2020b).

Then the goal is to design a loss function $\tilde{l}$ making the performance of the semantic segmentation model f trained on the noisy dataset $\widetilde{\mathcal{D}}$ is similar to the model performed on the clean dataset $\mathcal{D}$,

$$\tilde{f}^{*}_{\widetilde{\mathcal{D}},\tilde{l}} = \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} \tilde{l}\left(f\left(x_i\right), \tilde{y}_i\right). \tag{1}$$

2.2 TRANSITION MATRIX

It is noteworthy that the semantic segmentation model is trained on the noisy label $\widetilde{\mathcal{Y}}$ but validated on the clear label $\mathcal{Y}$. In this case, a transition matrix $T(x) \in [0,1]^{c \times c}$ is inevitable. We denote the $T(x)_{ij}$ as the transition probability of flipping the clean label $\mathcal{Y} = i$ to the noisy label $\widetilde{\mathcal{Y}} = j$, which is mathematically formulated as $T_{ij}(x) = P(\widetilde{\mathcal{Y}} = j \mid \mathcal{Y} = i, \mathcal{X} = x)$. Then clean class posterior can be inferred by the transition matrix $T(x)$ and noisy class posterior by means of the Bayesian inference, *i.e.*, $P(\mathcal{Y} \mid x) = T(x)P(\widetilde{\mathcal{Y}} \mid x)$. Generally, the transition matrix can be leveraged to combat noisy labels following three common ways. The first way is to utilize an adaptation layer that imitates the transition matrix in an end-to-end way to connect the clean labels and noisy labels. The second way is to correct the cross-entropy loss using the estimated transition matrix. The third way is to ease the model burden using the prior transition matrix.

2.3 TRANSITION MATRIX ESTIMATION

Anchor points method is widely used for estimating the transition matrix, where anchor points are some data samples belonging to certain semantic categories, which is mathematically denoted as $P\left(\mathcal{Y} = i \mid x^i\right) = 1, x^i \in \mathcal{X}$. Then we can acquire the transition matrix provided that anchor points exist and the label noise is instance-independent.

$$P\left(\widetilde{\mathcal{Y}} = j \mid x^i\right) = \sum_{k=1}^{C} P\left(\widetilde{\mathcal{Y}} = j \mid \mathcal{Y} = k, x^i\right) P\left(\mathcal{Y} = k \mid x^i\right) = P(\widetilde{\mathcal{Y}} = j \mid \mathcal{Y} = i, x) = T_{ij},$$
$$\text{s.t.} \quad P\left(\mathcal{Y} = k \mid x^i\right) = \begin{cases} 1, & k = i, \\ 0, & k \neq i. \end{cases} \tag{2}$$

According to the Equation (2), to estimate the transition matrix, it is necessary to find anchor points and estimate the noisy class posterior, and then the transition matrix is formulated as follows,

$$\hat{P}\left(\widetilde{\mathcal{Y}} = j \mid x^i\right) = \sum_{k=1}^{C} \hat{P}\left(\widetilde{\mathcal{Y}} = j \mid \mathcal{Y} = k, x^i\right) P\left(\mathcal{Y} = k \mid x^i\right) = \hat{P}(\widetilde{\mathcal{Y}} = j \mid Y = i, x) = \hat{P}_{ij}. \tag{3}$$

However, the anchor points are hard to identify by annotation algorithms during the estimation of the transition matrix. This de facto motivates us to consider an alternative way of avoiding approximating the transition matrix directly.

3 SEMANTIC ADAPTATION ESTIMATOR

To tackle the problem of semantic-dependent label noise caused by the anchor points, we resort to indirectly regressing the transition matrix by means of the proposed semantic adaptation estimator (SAE). Here we give a specific introduction in the following section.

3.1 SEMANTIC-DEPENDENT TRANSITION MATRICES

We formulate the transition matrix $T(x)$ using several introduced semantic-dependent transition matrices $T(x)_1, T(x)_2, \ldots, T(x)_N$, which can be found in Equation (4).

$$\begin{aligned} T(x) = P(\widetilde{\mathcal{Y}} = j \mid \mathcal{Y} = i) &= \sum_{l=1}^{N} \sum_{i=1}^{C} P\left(\widetilde{\mathcal{Y}}_l = j \mid \mathcal{Y}_l = l, \mathcal{Y} = i\right) P\left(\mathcal{Y}_l = l \mid \mathcal{Y} = i\right) \\ &= \sum_{l=1}^{N} \sum_{i=1}^{C_l} P\left(\widetilde{\mathcal{Y}}_l = j \mid \mathcal{Y}_l = l, \mathcal{Y} = i\right) = \sum_{l=1}^{N} T(x)_l, \end{aligned} \tag{4}$$

where $\widetilde{\mathcal{Y}}_l$ is the noisy labels for similar semantic categories group , $\mathcal{Y}_l$ is the clean labels for similar semantic categories group. C_l is the lth similar semantic categories group, where $C = \sum_{l=1}^{N} C_l$.

We assume that the probability of flipping any one semantic category in semantic-dependent transition matrix $T(x)_i$ to another semantic category in semantic-dependent transition matrix $T(x)_j$ is independent. Then we have $P\left(\mathcal{Y}_l = l \mid \mathcal{Y} = i\right) = 1$ for $i \in C_l$ and $P\left(\mathcal{Y}_l = l \mid \mathcal{Y} = i\right) = 0$, otherwise.

$$T(x)_l = \sum_{i=1}^{C_l} P\left(\widetilde{\mathcal{Y}}_l = j \mid \mathcal{Y}_l = l, \mathcal{Y} = i\right). \tag{5}$$

Then we formulate the semantic-dependent label noise problem related to the semantic-dependent transition matrices $T(x)_1, T(x)_2, \ldots, T(x)_N$ as follows,

$$\begin{aligned} \tilde{f}^{*}_{\widetilde{\mathcal{D}},\widetilde{l}} = \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} \tilde{l}\left(\sum_{l=1}^{N} T(x)_l f\left(x_i\right), \tilde{y}_i\right) &= \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} \tilde{l}\left(T(x)_1 f\left(x_i\right), \tilde{y}_i\right) \\ + \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} \tilde{l}\left(T(x)_2 f\left(x_i\right), \tilde{y}_i\right) + \ldots + \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} \tilde{l}\left(T(x)_N f\left(x_i\right), \tilde{y}_i\right). \end{aligned} \tag{6}$$

Intuitively, we decouple a complicated instance-dependent transition matrix to N sample semantic-dependent transition matrices shown in Equation (6). In such a manner, a hard problem can be decomposed into several simple sub-problems and conversely be solved by composing the corresponding sub-solutions. Nevertheless, inferring semantic-dependent transition matrices from N semantic category groups is still difficult. Based thereon, we in this work propose a semantic adaptation loss to partition the point cloud, where the proposed loss can be used to decouple similar semantic categories data and estimate each semantic-dependent transition matrices in the training phase.

3.2 SEMANTIC ADAPTATION LOSS

We modify the point cloud semantic segmentation backbone with an interpretable compositional layer Shen et al. (2021). Each filter in the encoder layer is assumed to stably decouple the similar semantic categories or the same region through different point cloud scenes. To ensure the similar semantic categories of the visual regions are decoupled by each filter, we design a group of filters

to represent the same point cloud region. The filters $\Omega = \{1, 2, \cdots, d\}$ in the encoder layer are organized by different independent groups of similar semantic categories $A_1, A_2, \cdots, A_N$, where $A_1 \cup A_2 \cup \cdots \cup A_N = \Omega$, $A_i \cap A_j = \emptyset$. $\mathbf{A} = \{A_1, A_2, \cdots, A_N\}$ is the combination of different filters. Given a series of point cloud scenes $\widetilde{\mathcal{P}}$, our goal is to decouple the training samples ensuring that filters in the same group consistently denote the same region through different scenes, and samples in different groups distribute different semantic adaptation estimator.

Given a point cloud data block $\widetilde{p}$, the $x_i^{\widetilde{p}}$ represents the feature map of the ith filter, and $X_i = \left\{x_i^{\widetilde{p}}\right\}_{\widetilde{p}\in\widetilde{\mathcal{P}}}$ represents a series of feature maps of the ith filter. Then, $\mathcal{K}$ is a kernel function, $s_{ij} = \mathcal{K}(X_i, X_j) \in \mathbb{R}$ measures the similarity between the ith and jth filter that represents the same region through different point cloud scenes. Then we propose a semantic adaptation loss to measure the differences between filters and train the filters as follows,

$$\mathrm{L}^{\text{semantic}}(\boldsymbol{\theta}, \mathbf{A}) = -\sum_{l=1}^{N} \frac{S_l^{\text{in}}}{S_l^{\text{all}}} = -\sum_{l=1}^{N} \frac{\sum_{i,j\in A_l} s_{ij}}{\sum_{i\in A_l, j\in\Omega} s_{ij}}, \tag{7}$$

where $\boldsymbol{\theta}$ is the parameters of the backbone, $S_l^{\text{in}} = \sum_{i,j\in A_l} s_{ij} = \sum_{i,j\in A_l} \mathcal{K}(X_i, X_j)$ represents the similarity between filters in the similar semantic group A_l, $S_l^{\text{all}} = \sum_{i\in A_l, j\in\Omega} s_{ij} = \sum_{i\in A_l, j\in\Omega} \mathcal{K}(X_i, X_j)$ represents the similarity between filters in A_l and all filters in Ω. Clearly, it is reasonable for the filters in the similar semantic group A_l have similar feature, which ensures that the same group consistently represent the same region through different scenes. Meanwhile, the part difference of these feature maps translates a variety of the same semantic of regions. Besides, the proposed loss makes filters in different groups have different feature maps, which ensures that the different semantic regions are decoupled by different groups of filters.

In addition, another loss is necessary to ensure that filters of the different semantic group A_l are able to respond to the same semantic of regions Shen et al. (2021). Given a point cloud data block with c semantic categories, we use $\widetilde{P}_l \in \widetilde{P}$ to represent the subset of point cloud of the semantic category l, $(l = 1, 2, \ldots, N)$. Note that the filters in a specific semantic category group will be activated by several specific semantic regions and kept silent on remained semantic regions. For this purpose, we design a quantification of distribution to represent each filter's activation on different categories. Given the pth point cloud data block, let $z_l^{(p)}$ represent the average activation score of filters in semantic group A_l, and we have $z_l^{(p)} = \frac{1}{|A_l|\cdot m} \sum_{i\in A_k} \sum_{u=1}^{m} x_{i,u}^{\widetilde{p}}$, where $|A_l|$ represents the number of filters in semantic group A_l, $x_{i,u}^{\widetilde{p}}$ represents the uth element in $x_i^{\widetilde{p}}$. $s_{pq} = \mathcal{K}\left(\mathbf{z}^{(p)}, \mathbf{z}^{(q)}\right) = \left(\mathbf{z}^{(p)}\right)^{\top}\left(\mathbf{z}^{(q)}\right) \in \mathbb{R}$ is the similarity of activation quantify distribution of different semantic groups on the pth and qth data block of point cloud. Next, we propose a semantic decouple loss to decouple the samples of different categories as follows,

$$\mathrm{L}^{\text{decouple}}(\boldsymbol{\theta}) = -\sum_{l=1}^{N} \frac{\sum_{p,q\in\widetilde{P}_l} s_{pq}}{\sum_{p\in\widetilde{P}_l, q\in\widetilde{P}} s_{pq}} = -\sum_{c=1}^{C} \frac{\sum_{p,q\in\widetilde{P}_l} \mathcal{K}\left(\mathbf{z}^{(p)}, \mathbf{z}^{(q)}\right)}{\sum_{p\in\widetilde{P}_l, q\in\widetilde{P}} \mathcal{K}\left(\mathbf{z}^{(p)}, \mathbf{z}^{(q)}\right)}. \tag{8}$$

Finally, the semantic adaptation loss is summarized as follows,

$$\mathrm{L}(\boldsymbol{\theta}, \mathbf{A}) = \lambda \mathrm{L}^{\text{semantic}}(\boldsymbol{\theta}, \mathbf{A}) + \beta \mathrm{L}^{\text{decouple}}(\boldsymbol{\theta}) + \mathrm{L}^{\text{seg}}, \tag{9}$$

where the coefficients λ, β are tuned during the training phase. We set λ = 0.1 and β = 0.1 in our work.

3.3 Anchor-free Estimator

As aforementioned, the anchor points are hard to be identified during the estimation of the transition matrix Xia et al. (2020). In this section, an anchor-free self-attention mechanism estimator is introduced for measuring the influence from one semantic category to another semantic categories. Specifically, the predicted semantic labels $\mathcal{Y} \in C \times 1$ performs matrix multiplication with its transposed form $\mathcal{Y}^T \in 1 \times C$. Then a semantic transition matrix is derived as follows Ren et al. (2022),

$$T(x)_{ij} = \frac{\exp\left(\mathcal{Y}_i \cdot \mathcal{Y}_j\right)}{\sum_{i=1}^{C}\left(\mathcal{Y}_i \cdot \mathcal{Y}_j\right)}, \tag{10}$$

where $T(x)_{ij}$ represents the probability of flipping the clean label $\mathcal{Y} = i$ to the noisy label $\widetilde{\mathcal{Y}} = j$.

Algorithm 1 Semantic Adaptation Estimator (SAE)

Input: Label noise training point cloud $\widetilde{\mathcal{D}}$.

1: Train a deep model according to Equation (1);
2: Minimize Equation (9) to decouple the training noisy label samples;
3: Use Equation (6) to decouple the instance-dependent transition matrix to semantic-dependent transition matrices;
4: Use Equation (10) to estimate the semantic-dependent transition matrices $T(x)_l$;
5: Couple the instance-dependent transition matrix for each semantic instance $T(x)_{ij}$ according to Use Equation (4);

Output: The estimated transition matrix $T(x)$.

3.4 Theoretical Analysis

In this section, we aim to provide a proof showing that the proposed SAE ensures statistical consistence of the optimal semantic segmentation within a reasonable assumption.

Assumption 1 The probability of flipping the semantic-dependent transition matrix $T(x)_i$ to another semantic-dependent transition matrix $T(x)_j$ is independent.

Theorem 1 *Under Assumption 1, the semantic segmentation model of the SAE is statistically consistent.*

Proof: We set the estimation error as $Error$, and we have:

$$\begin{aligned}
Error &= T(x)_{ij} - T(x)^{GT} = \sum_{l=1}^{N} T(x)_l - T(x)^{GT} = \sum_{l=1}^{N} T(x)_l - \sum_{u=1}^{N}\sum_{l=1}^{N} T(x)_{l,v}^{GT} \\
&= T(x)_1 + \ldots + T(x)_N - (T(x)_1^{GT} + \ldots + T(x)_N^{GT} + \sum_{l=1}^{N}\sum_{u=1,u\neq l}^{N} T(x)_{l,v}^{GT}) \\
&= (T(x)_1 - T(x)_1^{GT}) + \ldots + (T(x)_N - T(x)_N^{GT}) - \sum_{l=1}^{N}\sum_{u=1,u\neq l}^{N} T(x)_{l,v}^{GT} \\
&= \sum_{l=1}^{N}(T(x)_l - T(x)_l^{GT}) + C.
\end{aligned} \tag{11}$$

The Equation (11) holds because there is no additional estimation error for the transition matrices to decouple the transition from the instance-dependent to the semantic-dependent. $\sum_{l=1}^{N}\sum_{u=1,u\neq l}^{N} T(x)_{l,v}^{GT}$ is a constant. $T(x)_l - T(x)_l^{GT}$ is risk-consistent, and the corresponded estimator is therefore statistically consistent Song et al. (2022). Then, it is reasonable to learn the instance-dependent transition matrix by minimizing several statistically consistent semantic segmentation transition matrices because a favorable transition matrix should make the semantic segmentation risk small.

4 Experiments

In this section, we perform several experiments to demonstrate the performance of our method. First, we give an introduction of the datasets, and show the training details of our method. Then an group experiments are conducted to show the effectiveness in point cloud semantic segmentation tasks with various label noise levels. Furthermore, we provide the transition matrix estimation error results with different label noise levels of two datasets to show the correctness of the SAE.

4.1 EXPERIMENT SETUP

Datasets. We test the effectiveness of SAE on the corrupted version of two datasets, (*i.e.*, S3DIS Armeni et al. (2016), ScanNet Dai et al. (2017). S3DIS contains six large-scale indoor areas with 271 rooms. Following the same experimental setting as previous methods, Area1, Area2, Area3, Area4, and Area6 are used as the training set, and Area5 for testing. There are in total 12 semantic elements, which can be classified into window, door, and other structural elements. ScanNet is a popular 3D real-world dataset with label noise from human annotators, which contains 2.5 million views in 1513 scans. The first 1201 scenes are adopted as the training set, and the remaining are used for testing. We leverage the existing three annotation algorithms ((*i.e.*, Kpconv Thomas et al. (2019), MPRM Wei et al. (2020) and PCAM Wei et al. (2020)) to corrupt the training samples. Given that it is more realistic that annotation algorithms have no prior semantic knowledge, the loss function of the above annotation algorithms is vanilla cross entropy loss. In Step 1, we use the same training settings to train annotation algorithms as their papers are used. Taking Kpconv as an example, we train Kpconv base on Tensorflow and use an NVIDIA GeForce RTX 3090 graphics card. During the training phase, we use the batch size of 4 for S3DIS and ScanNet. In step 2, the training models are used to generate pseudo semantic labels. The final data format is (x, y, z, r, g, b, l), where l represents the semantic labels.

SAE implementation details and optimization In this work, we take a comparison with five state-of-the-art approaches (*i.e.* CE, MAE Ghosh et al. (2017), GCE Zhang & Sabuncu (2018) focal loss Lin et al. (2017), CGA Loss Lu et al. (2021)), and we use mean intersection over-union (mIoU) to evaluate the performance of each model on the clean test datasets. All experiments are implemented within the TensorFlow framework, which is trained for 500 epochs and tested both on NVIDIA Tesla V100. We first use SGD with momentum being 0.98, the batch size of 16, and an initial learning rate of 10^{-2} to initialize the network. During the training phase, $T(x)$ is initialized with an identity matrix. We use 16 filters in the interpretable compositional layer to decouple the datasets since we find that our method is insensitive the amount of filters and can obtain the highest mIoU with the number of filters being 16, which is shown in Figure. 2. Additionally, clean samples are not used for fine-tuning. And $T(x)$ should be removed during the test stage.

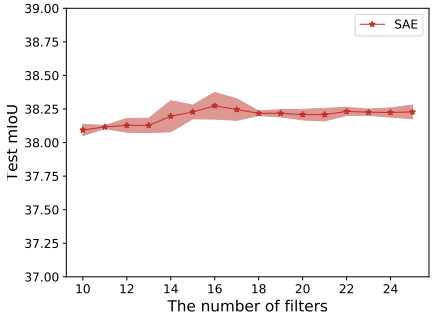

Figure 2: Illustration of the hyperparameter sensitivity. The figure illustrates how the number of filters affects the test semantic segmentation performance. The results (mean $\pm$ std) are reported over the last 5 epochs. The error bar for the standard deviation figure has been shaded.

4.2 SEMANTIC SEGMENTATION RESULTS EVALUATION

Table 1: Semantic segmentation mIoU (%) of several baselines methods on S3DIS with different label noise levels. The results (mean $\pm$ std) are reported over the last 10 epochs. The best results are boldfaced respectively.

	S3DIS		
	Kpconv	MPRM	PCAM
Forward(Patrini et al. (2017))	55.23±0.01	31.48±0.03	30.21±0.02
S-model(Goldberger & Ben-Reuven (2016))	55.14±0.02	32.73±0.04	31.04±0.01
Co-teaching(Han et al. (2018b))	52.37±0.02	39.13±0.02	27.16±0.02
M-correction(Arazo et al. (2019))	47.13±0.03	24.53±0.02	27.59±0.03
SAE	**64.48±0.13**	**38.29±0.06**	**33.96±0.05**

To further demonstrate the effectiveness of the SAE, five state-of-the-art approaches are implemented and tested with the default setting on different label noise levels in two datasets. We use the annotation algorithms' name to represent three semantic-dependent label noise levels (*i.e.*, Kp-

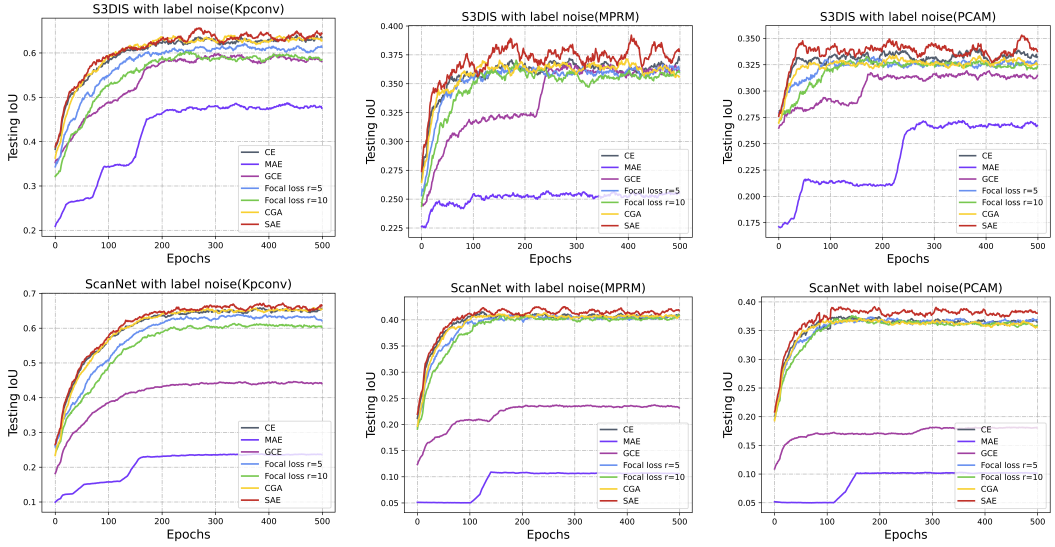

Figure 3: Illustration the effectiveness of SAE for point cloud semantic segmentation task with different label noise levels of two datasets.

conv is low-level label noise, MPRM is middle-level label noise, and PCAM is high-level label noise). The results are found in Table. 1 and Figure. 3. Obviously, our method improves the performance of corresponded backbone Kpconv in terms of all three semantic-dependent label noise levels in two datasets, which further demonstrates that our proposed method achieves consistent top results on label noise levels. We also observe that more interesting phenomena appear when we compare the robustness of different loss functions. MAE Ghosh et al. (2017) reports that MAE is much more robust than CE loss, however, which is not applicable to the semantic-dependent label noise in point cloud semantic segmentation task.

On the left of Table. 2, we give the specific semantic segmentation mIoU. It can be seen that our SAE for S3DIS with three label noise levels achieves 64.48±0.13%, 38.29±0.06%, 33.96±0.05%, respectively. As shown on the right of Table. 2, the performance on three label noise levels is significantly better than other algorithms. As a conclusion, these results show that the robust and stable of SAE is useful to the semantic segmentation on semantic-dependent label noise.

Table 2: Semantic segmentation mIoU (%) of state-of-the-art on S3DIS and ScanNet with different label noise levels. The results (mean ± std) are reported over the last 10 epochs. The best results are boldfaced respectively.

	S3DIS			ScanNet		
	Kpconv	MPRM	PCAM	Kpconv	MPRM	PCAM
CE	63.51±0.07	37.24±0.04	33.24±0.03	65.24±0.05	40.41±0.02	36.40±0.01
MAE	47.56±0.02	25.55±0.01	26.76±0.01	23.67±0.00	10.61±0.00	10.14±0.00
GCE	57.70±0.02	36.01±0.03	31.24±0.01	44.09±0.01	23.28±0.01	17.99±0.00
Focal loss r=5	61.17±0.07	36.12±0.03	32.45±0.01	61.85±0.06	40.46±0.00	36.80±0.07
Focal loss r=10	58.71±0.09	36.07±0.01	32.32±0.01	60.46±0.03	40.98±0.02	35.70±0.02
CGA	63.01±0.06	35.57±0.02	32.71±0.02	65.21±0.05	40.57±0.01	35.63±0.01
SAE	**64.48±0.13**	**38.29±0.06**	**33.96±0.05**	**66.30±0.08**	**41.23±0.04**	**37.99±0.02**

4.3 Transition Matrix Estimation Error

We compare the estimation error between our SAE and the pure estimator (E) on two real-world datasets (S3DIS Armeni et al. (2016) and ScanNet Dai et al. (2017)) under different testing epochs

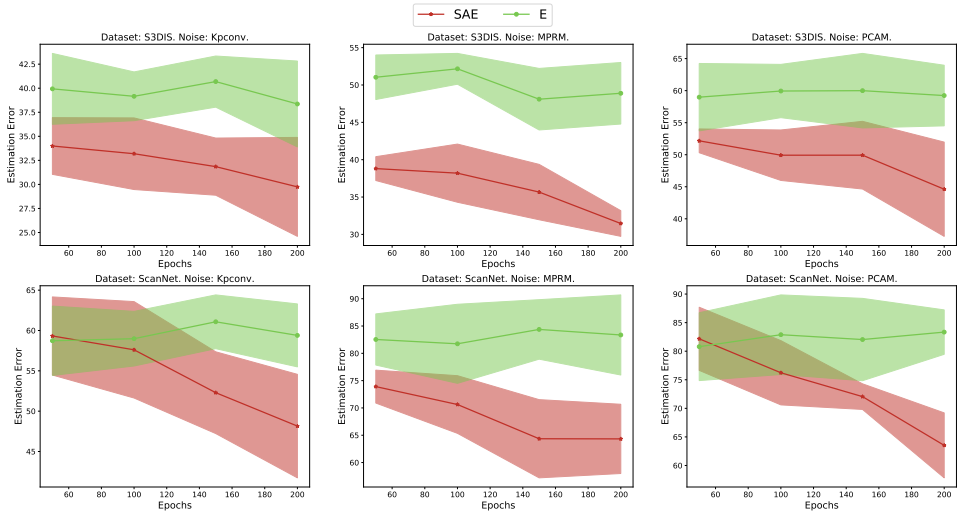

Figure 4: Transition matrix estimation error on S3DIS, ScanNet. The error bar for standard deviation in each figure has been shaded. The lower the better.

and different label noise levels. The network is trained for 500 epochs, and converges after 200 epochs. The estimation error is calculated by measuring the ℓ_1-distance between the estimated transition matrix and the ground truth $T(x)$. The average estimation error and the standard deviation over 5 repeated experiments for the both estimators are illustrated in Figure. 4.

Note that the SAE is an anchor-free estimation method which is different from the original $T(x)$ estimator Patrini et al. (2017). For a fair comparison, we use the pure anchor-free estimator (E) as the baseline. Figure. 4 illustrates the estimation error of the E and the SAE on two real-world datasets. Obviously, the estimation error of E keeps unchanged during the whole procedure. Consequently, the estimation error of the SAE is continuously smaller than the E. Moreover, the estimation error of the SAE is less sensitive to different label noise levels compared to the E. Besides, the estimation error of the E on MPRM noise is approximately doubled than the SAE on the convergence semantic segmentation model. In contrast, when testing the SAE with convergence models, its estimation errors on the different label noise levels show little difference. Similar to the results on the MPRM noise level, the experiments on another two label noise level also show that the estimation error of the SAE is continuously smaller than that of the E, which demonstrates the effectiveness of the proposed SAE. Notably, on the ScanNet dataset, both estimators perform a bit different in terms of estimation error, where the SAE surpasses the E after 100 epochs. Consequently, the proposed SAE method can correctly recover the transition matrix $T(x)$.

5 Conclusion

In this paper, we propose a new semantic adaptation estimator (SAE) to effectively model the semantic-dependent label noise in estimating the transition matrix $T(x)$. A divide-and-conquer paradigm is launched in our algorithm. Specifically, a semantic adaptation loss is used to decouple the complicated relationship of semantic categories. Then we estimate the relation among same regions. Finally, the coupled instance-dependent transition matrix is derived. Empirical evaluations on two real-world datasets (i.e., S3DIS and ScanNet) demonstrate the superior performance of our method, in comparison with the state-of-the-art. In the future, we are going to develop a contrast selection strategy to maximize the merit of automatic annotation algorithms. Meanwhile, we also will learn various prior constraints from different noise levels that in the same dataset. Overall, we expect our work will make appropriate contributions to diverse label noise applications.

REFERENCES

Ehsan Amid, Manfred K Warmuth, and Sriram Srinivasan. Two-temperature logistic regression based on the tsallis divergence. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2388–2396. PMLR, 2019a.

Ehsan Amid, Manfred KK Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on bregman divergences. *Advances in Neural Information Processing Systems*, 32, 2019b.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pp. 312–321. PMLR, 2019.

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1534–1543, 2016.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. 2016.

Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. *Advances in neural information processing systems*, 31, 2018a.

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. *Advances in neural information processing systems*, 31, 2018b.

Bo Han, Gang Niu, Xingrui Yu, Quanming Yao, Miao Xu, Ivor Tsang, and Masashi Sugiyama. Sigua: Forgetting may make learning with noisy labels more robust. In *International Conference on Machine Learning*, pp. 4006–4016. PMLR, 2020.

Qingyong Hu, Bo Yang, Guangchi Fang, Yulan Guo, Ales Leonardis, Niki Trigoni, and Andrew Markham. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds with 1000x fewer labels. *arXiv preprint arXiv:2104.04891*, 2021.

Qiangui Huang, Weiyue Wang, and Ulrich Neumann. Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2626–2635, 2018.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Jan Kremer, Fei Sha, and Christian Igel. Robust active label correction. In *International conference on artificial intelligence and statistics*, pp. 308–316. PMLR, 2018.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pp. 4313–4324. PMLR, 2020.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. volume 31, 2018.

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*, pp. 1910–1918, 2017.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pp. 6226–6236. PMLR, 2020.

Tao Lu, Limin Wang, and Gangshan Wu. Cga-net: Category guided aggregation for point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11693–11702, 2021.

Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. *arXiv preprint arXiv:1905.10045*, 2019.

Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *International Conference on Machine Learning*, pp. 3355–3364. PMLR, 2018.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*, pp. 6543–6553. PMLR, 2020.

Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *Advances in Neural Information Processing Systems*, 30, 2017.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. volume 30, 2017b.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Dayong Ren, Zhengyi Wu, Jiawei Li, Piaopiao Yu, Jie Guo, Mingqiang Wei, and Yanwen Guo. Point attention network for point cloud semantic segmentation. *Science China Information Sciences*, 65(9):1–14, 2022.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Wen Shen, Zhihua Wei, Shikun Huang, Binbin Zhang, Jiaqi Fan, Ping Zhao, and Quanshi Zhang. Interpretable compositional convolutional neural networks. *arXiv preprint arXiv:2107.04474*, 2021.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

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*, pp. 5552–5560, 2018.

Kiran K Thekumparampil, Ashish Khetan, Zinan Lin, and Sewoong Oh. Robustness of conditional gans to noisy labels. *Advances in neural information processing systems*, 31, 2018.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6411–6420, 2019.

Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 839–847, 2017.

Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2569–2578, 2018.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 322–330, 2019a.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5): 1–12, 2019b.

Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, and Lihua Xie. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4384–4393, 2020.

Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Nannan Wang, Haifeng Liu, and Gang Niu. Class2simi: A new perspective on learning with label noise. 2020.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33:7597–7610, 2020.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. *Advances in neural information processing systems*, 32, 2019.

Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *International Conference on Machine Learning*, pp. 10789–10798. PMLR, 2020a.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271, 2020b.

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*, pp. 7164–7173. PMLR, 2019.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 68–83, 2018.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.