Compensation Learning in Semantic Segmentation

Timo Kaiser, Christoph Reinders, Bodo Rosenhahn
Institute for Information Processing (tnt)
L3S - Leibniz Universität Hannover, Germany
{kaiser, reinders, rosenhahn}@tnt.uni-hannover.de

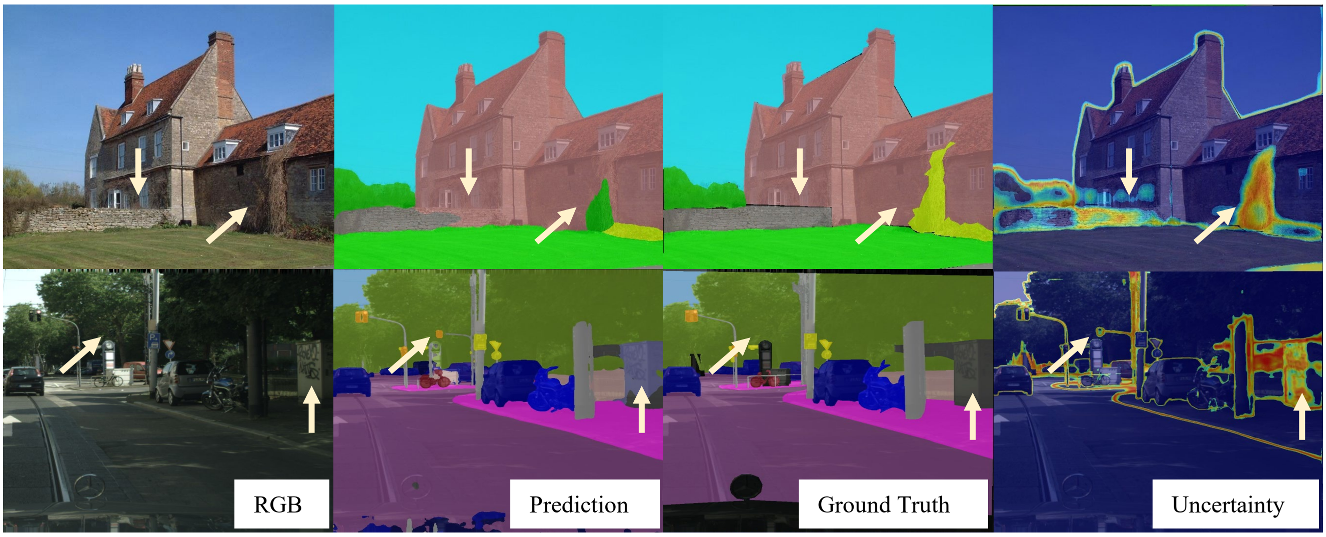

Figure 1. Our local compensation provides uncertainty estimation and localizes wrong model predictions and ground truth label noise in ADE20k (top) and Cityscapes (bottom). Arrows highlight regions of interest with errors and colors indicate different classes.

Abstract

Label noise and ambiguities between similar classes are challenging problems in developing new models and annotating new data for semantic segmentation. In this paper, we propose Compensation Learning in Semantic Segmentation, a framework to identify and compensate ambiguities as well as label noise. More specifically, we add a ground truth depending and globally learned bias to the classification logits and introduce a novel uncertainty branch for neural networks to induce the compensation bias only to relevant regions. Our method is employed into state-of-the-art segmentation frameworks and several experiments demonstrate that our proposed compensation learns inter-class relations that allow global identification of challenging ambiguities as well as the exact localization of subsequent label noise. Additionally, it enlarges robustness against label noise during training and allows target-oriented manipulation during inference. We evaluate the proposed method on Cityscapes, KITTI-STEP, ADE20k, and COCO-stuff10k.

1. Introduction

Semantic segmentation is a well-known and challenging task in computer vision [6, 34, 50]. Thanks to the large investment of time and resources, the research community published a large number of elaborately curated datasets to train and evaluate methods for semantic segmentation [16, 37, 53, 60, 79, 92]. Nevertheless, the industry needs an increasing amount of accurately annotated data and spends billion dollars to curate them [17]. Unfortunately, the annotation task stays challenging for humans even with advanced semi-automated annotation frameworks [1, 10, 72], because ambiguous image elements often can be assigned to multiple classes. Thus, annotated data is often noisy, with the consequence that the optimization of stochastic methods like neural networks is corrupted and the evaluation is distorted. Even the ground truth of widely used research benchmarks, which form the basis of this and many other papers, are subject to noise, as lamented by [42]. Semi-automated annotation without incorporating label noise is therefore a serious problem in semantic segmentation.

While tackling the impact of noisy labels is a well known research topic [7, 19, 31, 63], avoiding noisy labels during labeling is shallow investigated. Because modern semi-automated annotation frameworks estimate an initial guess with a pre-trained segmentation framework [1, 9, 10, 72], an obvious way to improve the annotation framework is to improve the segmentation framework. To remove the residual error in the estimate, the human curator is still asked to inspect and correct the entire image. To reduce this effort, uncertainty estimation can help to guide the curator to find the most likely error regions. Current approaches like *Bayesian Neural Networks* [52, 73], that estimate and incorporate uncertainty in semantic segmentation aim to make the training more robust against label noise, but mainly detect boundaries between neighboring segments [4, 7, 73].

Instead of using uncertainty estimation to make training more robust against noise, we aim to utilize robust training methods and uncertainty estimation to avoid new noise during data annotation. Therefore, we present a novel method transferring compensation learning to semantic segmentation to compensate noise and ambiguities with end-to-end trainable compensation weights. Compensation learning, which adds ground truth depending bias to model predictions, has been introduced by Yao *et al.* [85] for image classification. It allows the lowering of the influence of similar classes in order to reduce the impact of ambiguities and noise. We induce symmetry to make compensation learning stable during training and introduce an adaptive uncertainty branch that estimates the local importance of compensation.

Experiments on the widely used segmentation datasets Cityscapes [16], KITTI-STEP [79], ADE20k [92], and COCO-stuff10k [37] show that our method learns interpretable inter-class compensations and is able to estimate prediction uncertainties. We present how compensation identifies challenging class pairs and the uncertainty localizes prediction errors very accurately. Besides the interpretable guidance for data annotation, our method increases the robustness of training semantic segmentation methods with noisy labels and additionally introduces a useful method to improve the segmentation accuracy of certain classes. Moreover, we analyze and visualize inter-class ambiguities for the datasets.

In summary, our work contributes a novel framework[1] to improve semi-automated annotation that

- learns human-interpretable compensation weights of global inter-class ambiguities.
- introduces a novel uncertainty branch to adapt the compensation locally. The branch provides local guidance to image regions with high risk of errors.
- improves robustness against noise during training.
- allows application-oriented manipulation of segmentation accuracy during inference.

[1] Code available at https://github.com/tnt-LUH/compensation_learning

2. Related Work

Approaches to improve semi-automated annotation frameworks [1, 9, 10, 72] are stronger *Semantic Segmentation* methods to predict pseudo-labels, *Robust Learning* algorithms to reduce memorization of noise during fine-tuning, and *Uncertainty Estimation* as guidance to find prediction errors for human refinement. Related work for the mentioned topics and the state-of-the-art in *Robust Semantic Segmentation* are presented in this section.

Semantic Segmentation. The predominant approach for semantic segmentation methods is using convolutional neural networks with encoder/decoder [45] or feature pyramid [91] architectures. Extending the architecture with atrous convolutions [12] improves the accuracy and ended with the introduction of *DeepLabv3+* [13], which is widely used in science [14, 39, 74]. Latest state-of-the-art methods like *SegFormer* [83] apply transformers [56] to the architecture or use different approaches like Markov Random Fields [84], binary space partitioning [23], or class-agnostic clustering of associated pixels [89].

In general, the improvement of the above methods is accompanied by the introduction of improved backbones, such as [26, 27, 30, 44].

Robust Learning. Common methods handling label noise can be divided into label correction, loss correction methods, and meta-learning [63, 65, 75].

The goal of label correction is to identify and modify corrupted data annotations. Thereby, approaches vary from matching pseudo-labels to dynamic prototypes [88], estimating the non-affiliation to classes [33], or using bootstrapping, which maximizes the entropy between sample features and model predictions [2, 31, 58]. Bootstrapping is also used to approximate ensemble predictions [46]. Others directly optimize ground truth labels [41, 66, 76].

In loss correction approaches, the loss objective is adapted for each training sample. Whereas focal loss increases the impact of hard samples [36], others down-weight uncertain samples. Weights are obtained for example via mutual agreement of model ensembles [80] or *peer-predictions* [43, 93] and the lowest-k weighted samples are rejected [28]. Instead of weighting, other methods adapt the loss objective, *e.g.*, by combing loss functions [48], bounding losses [18, 90], or adding contrastive learning methods [87]. Assuming statistically consistent noise, a transition matrix models the probability of label flips between certain classes [70]. A known transition matrix helps to determine the clean prediction and it can be integrated into neural networks [21, 57, 64, 82, 86]. Some approaches learn transitions in an end-to-end manner [57, 64] or define it by human supervision [25]. Instead of modelling probabilities, Yao *et al.* [85] propose learning of ground truth depending bias. The influence of conditional noise can be compensated by adding bias to unconditional probability logits.

Meta-learning [71] with clear meta-data is used to predict additional information, *e.g.*, the expected noise per training sample to weight the loss [62] of prior predictions [65]. Meta-learning is also used to estimate the aforementioned transition matrix [78] or to augment new data by mixing meta- and noisy data [32, 35] or corrupting the meta-data [59].

Uncertainty Estimation. Uncertainty Estimation of neural networks is mandatory to evaluate automated decisions such as the creation of pseudo-labels during annotation. According to [20], approaches can be divided into *single deterministic methods* that predicts the uncertainty in one forward step [49, 54], *Bayesian methods* that utilize stochastic sampling [5, 22], *ensembles* to evaluate multiple predictions [24, 38, 69], and *test-time augmentation* [47].

Robust Semantic Segmentation. Many of the aforementioned methods are not applicable in semantic segmentation, either conceptually or in terms of complexity, or are applicable but not investigated further. Current state-of-the-art methods for robust learning in semantic segmentation detect label noise in an iterative process on the training set. Liu *et al.* [40] detect the memorization effect for every pixel and correct them with multi-scaled predictions. Since the model needs to be retrained from scratch after noise detection, this method cannot be reasonable applied in online semi-automated annotation. Wang *et al.* [77] propose a semi-supervised framework using contrastive predictive coding loss [55], but it does not identify label noise. Related to robust learning, uncertainty is incorporated in semantic segmentation methods. Atigh *et al.* [4] provide an uncertainty estimation by embedding semantic segmentation into hyperbolic space. Others estimate uncertainty with *Bayesian Neural Networks* [52, 73], model ensembles [29], or explicitly define uncertainty at region borders [7] or for entire segments via aggregated dispersion measures [61]. Although the latter go in the right direction, they do not explicitly consider ambiguities like our method.

3. Method

In this section, we present our proposed method of compensation learning in semantic segmentation that introduces global and local guidance for human label correction into existing segmentation frameworks. Furthermore, we introduce symmetry constraints that improve training and show how compensation can be used to manipulate inference of segmentation networks. The overall framework is presented in Fig. 2.

3.1. Preliminaries

The goal of semi-automated annotation tools is to accurately solve the semantic segmentation task with the least human curation effort. Semantic segmentation is a multi-class classification problem, in which each pixel x of an image I should be assigned to the true class label $y \in C$ from a set of classes $C = \{1, \ldots, K\}$. Modern annotation tools estimate an initial guess of the unknown label, which is then manually inspected and corrected to y by a human curator. The initial guess is nowadays estimated by neural networks F_Φ with optimized weights Φ (*e.g.*, by [13, 14, 51, 83]). During estimation, the probability $P(Y = i|x, \Phi)$ that represents the likelihood of pixel x belonging to class i is estimated for every $i \in C$. First, the classifier F_Φ predicts independent logits $\vec{l}_x = \{l_{x,1}, \ldots, l_{x,K}\}$, which are then transformed into conditional probabilities using the softmax function $\mathbb{S}$ [8]

$$P(Y = i|x, \Phi) = \mathbb{S}(\vec{l}_x)_i = \frac{e^{l_{x,i}}}{\sum_{n=1}^{K} e^{l_{x,n}}}, \qquad i \in C. \quad (1)$$

Finally, the pixel x gets assigned to class i with the highest probability.

Suitable weights Φ need to be obtained during a preceding optimization process with already annotated image data. The general approach is to minimize the *cross-entropy* loss

$$L_{\text{CE}} = -\frac{1}{|I|} \sum_{\substack{(x,\hat{y}) \\ x \in I}} \log\big(P(Y = \hat{y}|x, \Phi)\big), \quad \hat{y} \in C, \quad (2)$$

in which $\hat{y}$ denotes a given ground truth label for pixel x. Optionally, Φ can be fine-tuned on new annotated image data to decrease the domain gap and increase the segmentation accuracy for new estimations [79].

3.2. Global Compensation Learning

Unfortunately, optimizing the *cross-entropy* loss is prone to overfit on noisy or ambiguous pixels [81, 90]. The segmentation accuracy of the classifier F_Φ degrades during pretraining or induces confirmation bias during optional fine-tuning [3]. For noisy labels, the ground truth label $\hat{y}$ differs from the true label y. Most label flips $\hat{y} \neq y$ are based on ambiguities between $\hat{y}$ and y. For example, a curator might simply recognize a *bus* as a *truck* when it is in the distance. Thus, ambiguous visual appearance and label flips should be seen as conditional dependent.

An ambiguous pixel x that could equally be assigned to classes i or j usually has only one ground truth label, *e.g.*, $\hat{y} = j$. According to the visual similarity, a well trained classifier will generate similar logits $l_{x,i} \approx l_{x,\hat{y}}$, which cause approximately equal probabilities $P(Y = i|x, \Phi) \approx P(Y = \hat{y}|x, \Phi)$. This leads to a large loss in L_{CE}, even if the probability for the ground

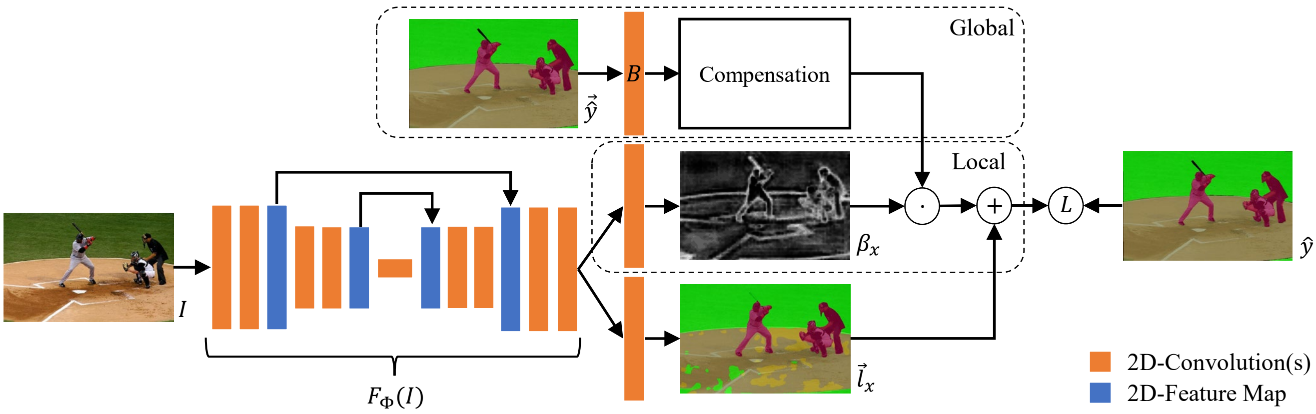

Figure 2. The global compensation B and local uncertainty β_x detects noisy image regions and makes convolutional segmentation frameworks $F_\Phi(I)$ robust against noise and ambiguities. It adds a ground truth $\hat{y}$ depending compensation to logits $\vec{l}_x$ to optimize L.

truth label $P(Y = \hat{y}|x, \Phi)$ has the highest probability and solves the classification task correctly. For the aforementioned example, the logit $l_{x,i}$ should therefore not strongly influence the probability $P(Y = \hat{y}|x)$. Technically, this intuition can be implemented by decreasing $l_{x,i}$ for pixels with the label $\hat{y} = j$.

To address ambiguities and label noise, we adapt compensation learning [85] and add a trainable conditional bias to the logits $\vec{l}_x$ during training. The softmax formulation from Eq. (1) is extended to

$$P(Y = i|x, \hat{y}, \Phi, B) = \frac{e^{l_{x,i}+B_{i\hat{y}}}}{\sum_{n=1}^{K} e^{l_{x,n}+B_{n\hat{y}}}}, \qquad i, \hat{y} \in C \tag{3}$$

where $B \in \mathbb{R}^{K\times K}$ denotes a compensation matrix. Each element $B_{i\hat{y}}$ enables decreasing or increasing the impact of logit $l_{x,i}$ for pixels with the ground truth label $\hat{y}$. Instead of manually defining B and to address two-dimensional images, we integrate B into the neural segmentation framework F_Φ with a single two-dimensional convolutional layer and optimize it alongside Φ. During training, compensation can reduce overfitting by minimizing Eq. (2) for ambiguous pixel regions and provide insights about inter-class correlations. The following sections describe how we improve compensation learning to boost semantic segmentation, and how we deduce a novel uncertainty estimation and inference approach.

3.3. Local Compensation Learning

The compensation matrix manipulates optimization globally so that an element $B_{i\hat{y}}$ affects the probabilities $P(Y = i|x, \hat{y}, \Phi)$ for all pixels with ground truth label $\hat{y}$. This is not reasonable for unambiguous pixels because it lowers the impact of good training samples in the optimization process. Thus, we introduce a novel uncertainty branch to estimate a local importance factor $\beta_x \in [0, 1]$ for each pixel x, and change the global concept of compensation Eq. (3) to

$$P(Y = i|x, \hat{y}, \Phi, B) = \frac{e^{l_{x,i}+\beta_x B_{i\hat{y}}}}{\sum_{n=1}^{K} e^{l_{x,n}+\beta_x B_{n\hat{y}}}}, \qquad i, \hat{y} \in C. \tag{4}$$

To predict the local importance β_x from high-level features, a lightweight regression head is added parallel to the classification head of the base segmentation framework. We employ two pointwise convolutional layers with 64 and 1 output channels, followed by a batchnorm and a sigmoid activation, respectively. The convolutional weights are added to Φ to be trained alongside the original segmentation framework. Our novel framework to observe conditional probabilities $\vec{p}_x$ for pixel x can be expressed with the corresponding logits $\vec{l}_x$, the softmax $\mathbb{S}$, and the one-hot ground truth vector $\vec{\hat{y}}$ as

$$\forall x \in I: \quad \vec{p}_x = \mathbb{S}(\vec{l}_x + \beta_x \cdot B\vec{\hat{y}}), \qquad \hat{y} \in C. \tag{5}$$

The general architecture is visualized in Fig. 2.

As ambiguities and label errors accompany, we propose compensation as local guidance to detect prediction errors during the annotation process. Inspired by the cost intensive uncertainty estimation of *Bayesian Neural Networks* [52], we introduce a lightweight approach to determine the uncertainty only with the uncompensated logits $\vec{l}_x$ and the global and local compensation B and β_x. Since the true annotation $y = \hat{y}$ is most likely in the top-k uncompensated predictions, we calculate the local variance for pixel x with

$$\sigma_x^2 = \frac{1}{k} \sum_{c \in C^k} \Big(\underbrace{P(Y = o|x, \hat{y} = c, \Phi)}_{\text{Eq. (4)}} - \underbrace{P(Y = o|x, \Phi)}_{\text{Eq. (1)}} \Big)^2, \tag{6}$$

where C^k denotes the subset of the top-k classes and o is the class with the highest uncompensated probability. The variance approaches zero, if β_x is small or the top-k classes do not compensate each other. In contrast, the variance increases for compensating classes in C^k, which indicates underlying ambiguities. To refine σ_x^2, we incorporate the model uncertainty u_x and estimate the likelihood of prediction errors e_x of pixel x with

$$\forall x \in I : e_x = \Bigg[\sigma_x^2 \cdot \underbrace{\left(1 - \max_{i \in C} P(Y = i | x, \Phi)\right)}_{u_x}\Bigg]^{\varphi} \in [0, 1] \tag{7}$$

and add an exponent φ that allows to amplify or diminish the uncertainty for visualization purposes. In our experiments, φ is adapted manually for visualization purposes only and we set k to 5 (see supplementary material, Sec. D).

3.4. Constrained and Penalized Compensation

The described compensation matrix B entails two risks: Adding a compensation matrix allows a mode collapse and can amplify the negative impact of imbalances in the training data.

Setting $\beta_x \approx 1$, $B_{i,i} \gg 0$ and $B_{i,j\neq i} \ll 0$ minimizes the main objective Eq. (2) without need of reasonable Φ. To prevent the mode collapse, we penalize B with local lasso regression [68] and extend the loss from Eq. (2) to

$$L = -\frac{1}{|I|} \sum_{\substack{(x,\hat{y}) \\ x \in I}} \log\Big(P(Y = \hat{y} | x, \Phi, B)\Big) + \frac{\alpha}{K} \sum_{i}^{C} \beta_x |B_{i\hat{y}}| \tag{8}$$

and weight the loss by α to adjust the penalty of B. We also constrain the diagonal entries of B to be zero:

$$\forall i \in C : \quad B_{ii} = 0\,, \tag{9}$$

as class i cannot have reasonable correlations to itself.

To enlarge the robustness against imbalances like the proportion of road pixels and sidewalk pixels, we add an optional symmetry constraint

$$\forall i, j \in C : \quad B_{ij} = B_{ji} \tag{10}$$

to stabilize the training. As drawback, this symmetry suppresses potential insights about the distribution of label errors from the global perspective as described in Sec. 3.2.

3.5. Compensated Inference

The described compensation framework relies on the ground truth label $\hat{y}$ and needs to be modified for the inference of unseen images for the semi-automated annotation task. A simple option is to remove B by setting Eq. (4) back to Eq. (1). For applications with prioritized classes, we instead propose to relax and estimate the ground truth one-hot vector $\vec{\hat{y}}$ (see Eq. (5)) with the uncompensated prediction of Eq. (1). While incorporating the model prediction, the usage of learned compensation B is not reasonable because it could induce confirmation bias for wrong predictions. Instead, the compensation matrix B can be defined manually to induce intended bias. Application-oriented compensations can prioritize important or vulnerable classes during prediction of the initial guess in the annotation process. To prioritize a class i in general, B_{ii} needs to be increased to a large positive value. To prioritize i against a specific class j, B_{ij} needs to be decreased to a large negative value, respectively. Our compensated inference can be applied a posteriori without the need of extra training and is therefore a non-bayesian alternative to [11].

4. Experiments

In this section, we present several experiments to evaluate the proposed method. First, the experimental setup and metrics used for evaluation are introduced. Then, we study how our method can be used to identify challenging inter-class ambiguities globally and prediction errors locally. Also the impact of our method on robustness against conditional label noise is evaluated and experiments are presented, which demonstrate the application-orientated usage of compensated inference.

4.1. Experimental Setup

We evaluate our method on the four publicly available semantic segmentation datasets Cityscapes [16], KITTI-STEP [79], ADE20k [92], and COCO-stuff10k [37]. These widely used datasets establish benchmarks for state-of-the-art segmentation methods with small and large amount of classes. Furthermore, the datasets Cityscapes and KITTI-STEP enable comparability for the interpretation task of ambiguities, because they share the equal set of classes C. Both datasets provide segmentation data in an autonomous driving setting, where the images are annotated per-pixel with 19 semantic classes, whereas ADE20k is annotated with 151 and COCO with 170 classes. We evaluate on the validation sets to allow extensive investigations.

To analyze the impact of compensation, we integrate our method into the well-known semantic segmentation framework *DeepLabv3+* [13] and the state-of-the-art framework *SegFormer* [83]. The loss balancing α is set to 0.01 and 1, the learning rate to 0.01 and 0.00006, and we train for 80 000 and 160 000 epochs, respectively.

Ours and reported reference methods are employed on top of the baseline frameworks and implemented in the widely used framework *MMSegmentation* [15] to improve the reproducability. More details on the experimental setup and implementations of later mentioned reference methods can be found in the supplementary material, Sec. A.

To evaluate our method, we use the widely-used mIoU metric [67] that evaluates the assignment of class labels and balances underrepresented classes. We also use the class accuracy (Acc_c) and aggregated accuracy (Acc_a) to verify segmentation results on pixel level. Mathematical details

Table 1. The learned compensation values B_{ij} for 11 classes in KITTI-STEP and Cityscapes provided by our method. A negative value B_{ij} lowers the impact of class i for pixels with the ground truth annotation j. Note that values are rounded after the first digit.

KITTI-STEP B_{ij}	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*tr. light*	*tr. sign*	*vegetation*	*terrain*	*sky*
road	0	-1.7	0	0	-0.2	-0.2	0	-0.1	-0.3	-1.1	0
sidewalk	-1.6	0	-0.4	-0.1	-0.1	-0.2	0	0	-0.4	-0.8	0
building	0	-0.4	0	-0.1	-0.3	-0.8	-0.1	-0.2	-1.7	-0.1	-0.5
wall	0	0	-0.1	0	-0.2	0	0	0	-0.1	0	0
fence	-0.1	-0.1	-0.2	-0.2	0	-0.1	0	0	-0.7	-0.3	0
pole	-0.1	-0.2	-0.8	0	-0.1	0	-0.1	-0.2	-1.1	-0.2	-0.3
tr. light	0	0	0	0	0	-0.1	0	0	-0.1	0	0
tr. sign	0	0	-0.2	0	0	-0.2	0	0	-0.4	0	0
vegetation	-0.3	-0.5	-1.8	-0.2	-0.8	-1.2	-0.2	-0.5	0	-1.3	-1.2
terrain	-1.0	-0.7	0	0	-0.3	-0.2	0	0	-1.2	0	0
sky	0	0	-0.4	0	0	-0.3	0	-0.1	-1.1	0	0

Cityscapes B_{ij}	*road*	*sidewalk*	*building*	*wall*	*fence*	*pole*	*tr. light*	*tr. sign*	*vegetation*	*terrain*	*sky*
road	0	-2.1	-0.1	-0.1	-0.1	-0.1	0	0	-0.2	-0.6	0
sidewalk	-2.0	0	-0.8	-0.3	-0.3	-0.5	0	0	-0.4	-1.0	0
building	-0.1	-0.8	0	-0.9	-0.9	-1.9	-0.4	-0.8	-2.5	-0.2	-0.8
wall	0	-0.3	-0.8	0	-0.6	-0.1	0	0	-0.5	-0.1	0
fence	0	-0.2	-0.7	-0.6	0	-0.3	0	0	-0.6	-0.2	0
pole	0	-0.4	-1.7	-0.1	-0.3	0	-0.2	-0.3	-1.3	-0.2	-0.2
tr. light	0	0	-0.3	0	0	-0.2	0	0	-0.3	0	0
tr. sign	0	0	-0.7	0	0	-0.3	-0.1	0	-0.3	0	0
vegetation	-0.1	-0.5	-2.4	-0.6	-0.7	-1.4	-0.3	-0.4	0	-1.1	-0.9
terrain	-0.4	-1.0	-0.1	-0.1	-0.2	-0.2	0	0	-1.0	0	0
sky	0	0	-0.7	0	0	-0.2	0	0	-0.8	0	0

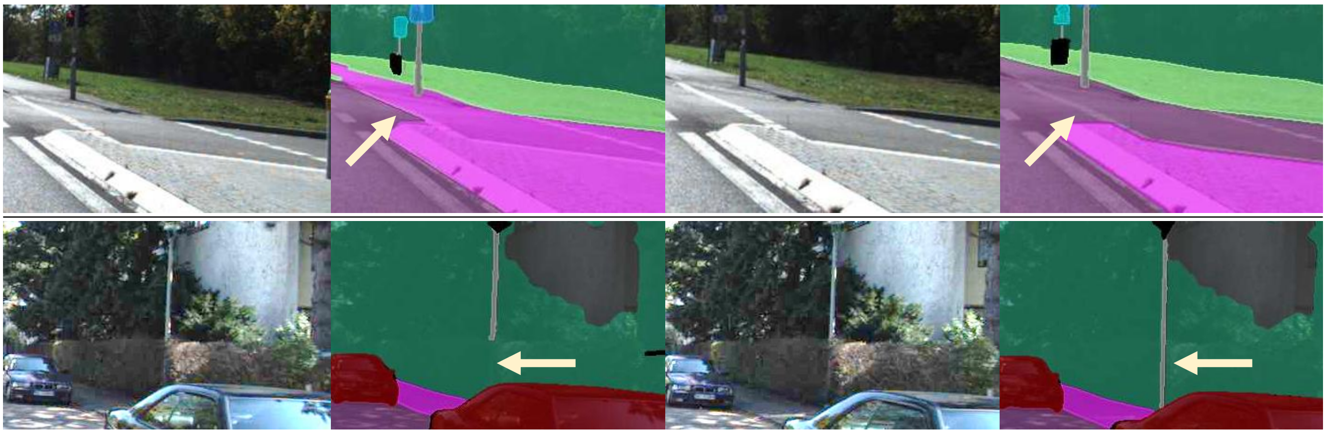

Figure 3. Most challenging class pairs indicated by learned compensation weights. The first and third column show subsequent images from KITTI-STEP video sequences and their ground truth. The upper row shows a flip between *road* and *sidewalk*, and the lower between *pole* and *vegetation*. Both ambiguous pairs are challenging and cause systematical errors.

of the metrics are elaborated in the supplementary material, Sec. B.

4.2. Global Compensation: Interpretable Data

In this section, we analyze the learned compensation weights B after training with Eq. (5) but without symmetry constraint (Eq. (10)). Tab. 1 shows a subselection of B trained on the low class datasets KITTI-STEP and Cityscapes with *DeepLabv3+*. In direct comparison, both datasets share the strongest compensated class pairs, *e.g.*, *road-sidewalk* or *pole-building*. The outstanding class pairs ij with large compensation weights B_{ij} can be considered more difficult for the curator and prone to label noise. Our method identifies intuitive, like *road-sidewalk*, and also not intuitive ambiguities, such as *pole-building* and *pole-vegetation*. These ambiguities can be verified by multiple samples of label flips in the data. Examples for the most ambiguous class pairs are shown in Fig. 3. Compared to the distribution dependent confusion matrix, the compensation matrix indicates ambiguities independently based on their impact during optimization. The full compensation matrices for *SegFormer* and *DeepLabv3* and confusion matrices can be found in the supplementary material, Sec. G and H. We noticed, that compensation stronger influences the transformer based *SegFormer*, because values in B are much larger compared to *DeepLabv3+*.

In summary, the proposed method learns and provides human-interpretable insights about inter-class ambiguities in the model optimization. With this information, human curators are able to improve dataset quality by focusing on systematical errors caused by ambiguities.

4.3. Local Compensation: Label Noise Detection

In the next experiment, we evaluate the ability of local compensation to locate potential prediction errors in the label estimation step during the annotation process. To verify the ability of noise detection, we mime the label correction process on the validation datasets by replacing the ratio

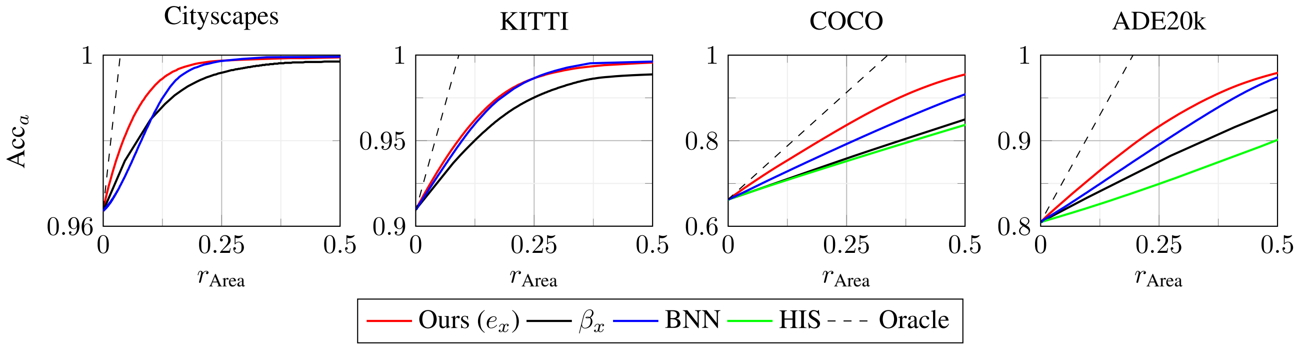

Figure 4. Analysis of error detection guided by our method used to correct prediction errors of *DeepLabv3+*. The ratio of top-k uncertain pixels r_{Area} that is corrected by the human curator improves the segmentation accuracy Acc_a. We report the accuracy after correction with our guidance e_x, the local compensation weight β_x, BNN [52], and HIS [4] (if available). Oracle mimes the optimal uncertainty estimation.

Table 2. Area-under-Curve for the noise detection experiments in Fig. 4 of our method compared to BNN [52] and HIS [4]. Note that HIS evaluation is only available for the large-scale datasets.

	AUC↑	Cityscapes	KITTI	COCO	ADE20K
	Oracle	0.999	0.996	0.943	0.981
Method	β_x	0.995	0.980	0.845	0.926
	BNN [52]	0.996	0.986	0.881	0.944
	HIS [4]	-	-	0.835	0.902
	Ours (e_x)	**0.997**	**0.987**	**0.909**	**0.951**

r_{Area} of the most uncertain pixels of the model prediction with ground truth data. We estimate our uncertainty e_x from Eq. (7) and compare it with the naïve β_x, a *Bayesian Neural Network* (BNN) approach [52], and with the hyperbolic segmentation framework (HIS) from Atigh *et al.* [4] for the large-scale datasets. With these uncertainties, we evaluate the accuracy with increasing r_{Area}. The results are shown in Fig. 4 and the corresponding area-under-curve metrics are provided in Tab. 2. The refined uncertainty estimation e_x boosts β_x and provides trustworthy guidance for label correction. For example, our method indicates approx. 50% of all prediction errors on Cityscapes with high precision by selecting less than 5% of the area. Our lightweight approach continuously outperforms the computational expensive BNN as well as the hyperbolic approach HIS. Note that the annotator is interested in high accuracy for small r_{Area}, since the area needs to be inspected manually.

Image samples and the corresponding e_x, model predictions and ground truth data are visualized in Fig. 1. It shows the fine-grained uncertainty prediction and how e_x can be used to identify prediction errors. Interestingly, we found multiple inconspicuous errors in the ground truth data, which are predicted correctly by our trained models (see the highlighted traffic-light). More samples and more detailed figures are provided in the supp. material, Sec. E.

Table 3. Mean intersection over union of our method compared to the baseline frameworks *DeepLabv3+* and *SegFormer* and applied robust learning methods *LogComp*, *s-model*, and *c-model*. Note that *c-model* is not applicable to large-scale experiments.

mIoU ↑	Cityscapes	KITTI	COCO	ADE20K	Mean
DeepLabv3+	0.797	0.570	**0.358**	**0.431**	0.539
+ *LogComp* [85]	0.783	0.567	0.356	**0.431**	0.534
+ *s-model* [21]	**0.799**	0.569	0.329	0.315	0.503
+ *c-model* [21]	0.219	0.137	-	-	-
+ **Ours**	0.799	**0.574**	0.354	0.429	0.539
+ **Ours (+ Sym)**	0.798	0.572	**0.358**	**0.431**	**0.541**
SegFormer	**0.821**	0.625	0.413	0.482	0.585
+ *LogComp* [85]	0.785	0.628	0.427	0.471	0.578
+ *s-model* [21]	**0.821**	0.570	0.430	0.482	0.576
+ *c-model* [21]	0.492	0.489	-	-	-
+ **Ours**	0.816	0.649	0.428	0.459	0.596
+ **Ours (+ Sym)**	**0.821**	**0.658**	**0.432**	**0.485**	**0.599**

Table 4. Average prediction uncertainty with and without compensation learning in the baseline framework *DeepLabv3+* indicating overfitting, *a.k.a.* memorization effect [40]. Note that we removed all compensation related weights during inference.

u_x ↑	Compensation	Cityscapes	KITTI	COCO	ADE20K
Train	✗	0.027	0.028	0.076	0.100
	✓	**0.048**	**0.035**	**0.081**	**0.122**
Val	✗	0.035	0.039	0.199	0.141
	✓	**0.058**	**0.048**	**0.204**	**0.160**

4.4. Robust Compensation: Training with Noise

Compensation learning decreases the impact of ambiguities on the loss during training and therefore enlarges the robustness against ambiguity based label noise and overfitting. To evaluate the impact of our method on the robustness, we first evaluate the impact of our method with and without the optional symmetry constraint on the segmentation performance of *SegFormer* and *DeepLabv3+* (see Tab. 3). To compare the results, we evaluate the simple *s-model* and complex *c-model* transition matrix approach as described in [21]. While *s-model* and *c-model* show instabilities in either large and/or small scale datasets, our

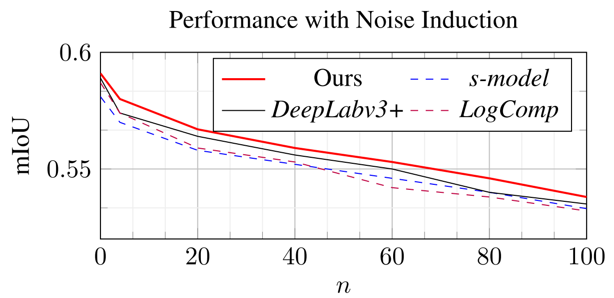

Figure 5. Performance degradation of segmentation quality after induction of label flips with noise parameter n showing the impact of our method on the robustness of *DeepLabv3+* compared to *LogComp* and *s-model*.

method achieves the same or higher mIoU than the baseline. We also evaluate the naïve and unconstrained compensation method *LogComp* as proposed by Yao *et al.* [85] and show that our contributions are mandatory to apply compensation learning in semantic segmentation. Especially the transformer based *SegFormer* benefits on the small dataset KITTI-STEP and increases the mIoU by 3.3 percent points.

Moreover, we induce human-like label noise between similar classes and corrupt the ground truth data during training. Following the approach of Liu *et al.* [40], we dilate the area of predefined classes with neighboured pixels of similar appearance that have a distance of at most n pixels. Comparable noise patterns can be found in the ground truth data of our datasets. Fig. 5 shows the accuracy degradation after inducing different label noise levels. While *LogComp* and *s-model* decrease the accuracy of the baseline, our method continuously improves the performance. Note that we restrict this expensive experiment to *DeepLabv3+* with KITTI-STEP and reduce the batchsize to 4 for ecological reasons. For detailed information about noise induction and visual samples, see supplementary material, Sec. C.

To show the influence of compensation against overfitting, we measure the average model uncertainty u_x of *DeepLabv3+* trained with and without compensation. Because overfitting manifests in certain predictions for ambiguous or noisy labels, also known as memorization [40], a robust model should stay uncertain for those uncertain regions even after long training. The comparison of the average model uncertainty u_x is given in Tab. 4. On validation and training data, the model with our proposed method is significantly more uncertain with a factor up to 1.7.

Overall, the experiments show that our method is able to learn interpretable guidance for label correction while improving robustness against noise by avoiding memorization.

4.5. Compensated Inference: Bias Induction

To outline the possibilities of induced compensation during inference as explained in Sec. 3.5, we present an exemplary application. In KITTI-STEP, vulnerable classes like

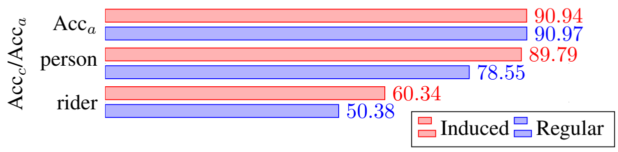

Figure 6. Induction Experiments on KITTI-STEP.

rider or *person* are expected to be more important for applications in autonomous driving. Therefore, we amplify the segmentation likelihood of those classes during inference by manually defining elements in the compensation matrix B. The exact value of the elements is determined experimentally. We set B_{ii} with $i \in \{\textit{person}, \textit{rider}\}$ to 30 and B_{ij} with $j \in \{\textit{sidewalk}, \textit{building}\}$ to -8. The resulting accuracy of the vulnerable classes is compared against the unmodified model in Fig. 6. Without loosing notable accumulated accuracy, the accuracy of our selected vulnerable classes increases by approx. 10 percent points. This shows the general ability to manipulate inference with compensations to improve annotation for given tasks with prioritized classes. The application applied on Cityscapes and all class accuracy metrics can be found in the supplementary material, Sec. F. Note that this experiment briefly outlines future possibilities, but rules to determine exact values in B need to be further investigated.

5. Conclusion

In this paper, we present compensation learning in semantic segmentation, a lightweight approach to learn and visualize inter-class relations to tackle ambiguity based label noise during the annotation of new semantic segmentation datasets. Our method creates a ground truth depending bias to compensate the influence of similar classes and ambiguities. The experiments demonstrate that our compensation learning method provides global and local guidance in the label correction process and introduces a powerful uncertainty estimation approach. Moreover, it improves the robustness against conditional label noise and accurately detects prediction errors of segmentation networks. We have presented insights about challenging class pairs in the Cityscapes, KITTI-STEP, ADE20k, and COCO datasets. This contribution helps the community to make semantic segmentation more robust against inter-class ambiguities and subsequent label noise.

6. Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor (grant no. 01DD20003) and the AI service center KISSKI (grant no. 01IS22093C), the Center for Digital Innovations (ZDIN) and the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122).

References

[1] Mykhaylo Andriluka, Jasper R. R. Uijlings, and Vittorio Ferrari. Fluid annotation: A human-machine collaboration interface for full image annotation. In *Proceedings of the 26th ACM International Conference on Multimedia*, 2018. 1, 2

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 2

[3] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks*, 2020. 3

[4] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 2, 3, 7

[5] David Barber and Christopher M Bishop. Ensemble learning in bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 1998. 3

[6] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[7] Patrik Olã Bressan, José Marcato Junior, José Augusto Correa Martins, Maximilian Jaderson de Melo, Diogo Nunes Gonçalves, Daniel Matte Freitas, Ana Paula Marques Ramos, Michelle Taís Garcia Furuya, Lucas Prado Osco, Jonathan de Andrade Silva, et al. Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. *International Journal of Applied Earth Observation and Geoinformation*, 2022. 2, 3

[8] John Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in Neural Information Processing Systems*. Morgan-Kaufmann, 1990. 3

[9] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixe, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 2

[10] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 1, 2

[11] Robin Chan, Matthias Rottmann, Radin Dardashti, Fabian Huger, Peter Schlicht, and Hanno Gottschalk. The ethical dilemma when (not) setting up cost-based decision rules in semantic segmentation. In *CVPRW*, 2019. 5

[12] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2

[13] 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*, 2018. 2, 3, 5

[14] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3

[15] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020. 5

[16] 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. 1, 2, 5

[17] S Feldman. The people and process behind artificial intelligence. Statista, 2019. 1

[18] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020. 2

[19] Benoit Frenay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2014. 2

[20] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021. 3

[21] Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *Proceedings of the International Conference on Learning Representations*, 2017. 2, 7

[22] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 2011. 3

[23] Daniel Gritzner and Jörn Ostermann. Semantic segmentation of aerial images using binary space partitioning. In *KI 2021: Advances in Artificial Intelligence*, 2021. 2

[24] Huaping Guo, Hongbing Liu, Ran Li, Changan Wu, Yibo Guo, and Mingliang Xu. Margin & diversity based ordering ensemble pruning. *Neurocomputing*, 2018. 3

[25] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. In *Advances in Neural Information Processing Systems*, 2018. 3

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 2

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 2

[28] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*, 2018. 2

[29] Christopher J. Holder and Muhammad Shafique. Efficient uncertainty estimation in semantic segmentation via distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3087–3094, October 2021. 3

[30] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *The IEEE International Conference on Computer Vision*, October 2019. 2

[31] Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems*, 2020. 2

[32] Timo Kaiser, Lukas Ehmann, Christoph Reinders, and Bodo Rosenhahn. Blind knowledge distillation for robust image classification, 2022. 3

[33] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

[34] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[35] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. 3

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Proceedings of the European Conference on Computer Vision*, 2014. 1, 2, 5

[38] Jakob Lindqvist, Amanda Olmin, Fredrik Lindsten, and Lennart Svensson. A general framework for ensemble distribution distillation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing*. IEEE, 2020. 3

[39] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L. Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019. 2

[40] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive early-learning correction for segmentation from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 3, 7, 8

[41] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. *arXiv preprint arXiv:2202.14026*, 2022. 2

[42] Xinyu Liu, Wuyang Li, Qiushi Yang, Baopu Li, and Yixuan Yuan. Towards robust adaptive object detection under noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 1

[43] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021. 2

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[46] Yangdi Lu and Wenbo He. Selc: Self-ensemble label correction improves learning with noisy labels. *IJCAI*, 2022. 2

[47] Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, Dmitry Molchanov, and Dmitry Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020. 3

[48] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2

[49] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 2018. 3

[50] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[51] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 2021. 3

[52] Jishnu Mukhoti and Yarin Gal. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv preprint arXiv:1811.12709*, 2018. 2, 3, 4, 7

[53] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[54] Philipp Oberdiek, Matthias Rottmann, and Hanno Gottschalk. Classification uncertainty of deep neural networks based on gradient information. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, 2018. 3

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[56] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 10–15 Jul 2018. 2

[57] 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*, 2017. 2

[58] 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. 2

[59] Christoph Reinders, Frederik Schubert, and Bodo Rosenhahn. Chimeramix: Image classification on small datasets via masked feature mixing. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1298–1305. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track. 3

[60] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision*, 2016. 1

[61] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *IJCNN*, 2020. 3

[62] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 2019. 3

[63] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020. 2

[64] Sainbayar Sukhbaatar and Rob Fergus. Learning from noisy labels with deep neural networks. In *International Conference on Learning Representations, Workshop Track Proceedings*, 2015. 2

[65] Haoliang Sun, Chenhui Guo, Qi Wei, Zhongyi Han, and Yilong Yin. Learning to rectify for robust learning with noisy labels. *Pattern Recognition*, 2022. 2, 3

[66] 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. 2

[67] Martin Thoma. A survey of semantic segmentation. *arXiv preprint arXiv:1602.06541*, 2016. 5

[68] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996. 5

[69] Matias Valdenegro-Toro. Deep sub-ensembles for fast uncertainty estimation in image classification. *arXiv preprint arXiv:1910.08168*, 2019. 3

[70] Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 2018. 2

[71] Joaquin Vanschoren. Meta-learning. In *Automated Machine Learning*. Springer, Cham, 2019. 3

[72] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[73] Jianfeng Wang and Thomas Lukasiewicz. Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 2, 3

[74] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[75] 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*, 2019. 2

[76] Yikai Wang, Xinwei Sun, and Yanwei Fu. Scalable penalized regression for noise detection in learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 2

[77] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 3

[78] Zhen Wang, Guosheng Hu, and Qinghua Hu. Training noise-robust deep neural networks via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[79] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljosa Osep, Laura Leal-Taixe, and Liang-Chieh Chen. Step: Segmenting and tracking every pixel. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 1, 2, 3, 5

[80] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[81] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *In-*

ternational Conference on Learning Representations, 2022. 3

[82] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems*, 2019. 2

[83] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021. 2, 3, 5

[84] Zhiwei Xu, Thalaiyasingam Ajanthan, and Richard Hartley. Fast and differentiable message passing on pairwise markov random fields. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[85] Rujing Yao, Ou Wu, and Mengyang Li. Compensation learning. *arXiv preprint arXiv:2107.11921*, 2021. 2, 3, 4, 7, 8

[86] 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*, 2020. 2

[87] Li Yi, Sheng Liu, Qi She, A. Ian McLeod, and Boyu Wang. On learning contrastive representations for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022. 2

[88] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[89] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2021. 2

[90] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018. 2, 3

[91] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[92] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5

[93] Zhaowei Zhu, Tongliang Liu, and Yang Liu. A second-order approach to learning with instance-dependent label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2