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

RGB Prediction Ground Truth

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.

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.

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.
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.

\(^1\)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 predict 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 \( F_\Phi \) with optimized weights \( \Phi \) (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 \( F_\Phi \) predicts independent logits \( l^i_x = \{l_{x,1}, \ldots, l_{x,K}\} \), which are then transformed into conditional probabilities using the softmax function \( \Sigma \) [8]:

\[
P(Y = i|x, \Phi) = \Sigma(\Phi)_i = \frac{e^{l_{x,i}}}{\sum_{n=1}^{K} e^{l_{x,n}}}, \quad i \in C. \tag{1}
\]

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

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

\[
L_{CE} = -\frac{1}{|I|} \sum_{(x, \hat{y}) \in I} \log \left(P(Y = \hat{y}|x, \Phi)\right), \quad \hat{y} \in C, \tag{2}
\]

in which \( \hat{y} \) denotes a given ground truth label for pixel \( x \). Optionally, \( \Phi \) 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 \( F_\Phi \) degrades during pre-training 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 \( l_{x,i} \approx l_{x,j} \), which cause approximately equal probabilities \( P(Y = i|x, \Phi) \approx P(Y = j|x, \Phi) \). This leads to a large loss in \( L_{CE} \), even if the probability for the ground
truth label $P(Y = y|x, \Phi)$ 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 $P(Y = y|x)$. Technically, this intuition can be implemented by decreasing $l_{x,i}$ for pixels with the label $y = j$.

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

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

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 $F_{\Phi}$ with a single two-dimensional convolutional layer and optimize it alongside $\Phi$. 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

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

(4)

To predict the local importance $\beta_x$ 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 $\Phi$ to be trained alongside the original segmentation framework. Our novel framework to observe conditional probabilities $\hat{p}_x$ for pixel $x$ can be expressed with the corresponding logits $l_x$, the softmax $S$, and the one-hot ground truth vector $\hat{y}$ as

$$\forall x \in I : \quad \hat{p}_x = S(l_x + \beta_x \cdot B \hat{y}), \quad \hat{y} \in C.$$

(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 $l_x$ and the local and global compensation $B$ and $\beta_x$. 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

$$\sigma_x^2 = \frac{1}{k} \sum_{c \in C^k} \left( \frac{P(Y = o|x, y = c, \Phi) - P(Y = o|x, \Phi)}{\text{Eq. (4)}} \right)^2,$$

(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 $\beta_x$ 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

Figure 2. The global compensation $B$ and local uncertainty $\beta_x$ 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 $l_x$ to optimize $L$. 

3269
underlying ambiguities. To refine $\sigma_x^2$, we incorporate the model uncertainty $u_x$ and estimate the likelihood of prediction errors $e_x$ of pixel $x$ with

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

(7)

and add an exponent $\varphi$ that allows to amplify or diminish the uncertainty for visualization purposes. In our experiments, $\varphi$ 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 $\Phi$. To prevent the mode collapse, we penalize $B$ with local lasso regression [68] and extend the loss from Eq. (2) to

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

(8)

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

$$
\forall i \in C : B_{i,i} = 0,
$$

(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

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

(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 $\hat{y}$ (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-oriented 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 $\alpha$ 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$_{Agg}$) 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.

<table>
<thead>
<tr>
<th>$B_{ij}$</th>
<th>KITTI-STEP</th>
<th>Cityscapes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>road</td>
<td>sidewalk</td>
</tr>
<tr>
<td>road</td>
<td>0</td>
<td>-0.3</td>
</tr>
<tr>
<td>sidewalk</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>building</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>wall</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fence</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>pole</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>tr. light</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tr. sign</td>
<td>0</td>
<td>-0.2</td>
</tr>
<tr>
<td>vegetation</td>
<td>0.3</td>
<td>-0.5</td>
</tr>
<tr>
<td>terrain</td>
<td>0.1</td>
<td>-0.7</td>
</tr>
<tr>
<td>sky</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 sub-selection 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...
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cityscapes</th>
<th>KITTI</th>
<th>COCO</th>
<th>ADE20K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>0.999</td>
<td>0.996</td>
<td>0.943</td>
<td>0.981</td>
</tr>
<tr>
<td>BNN [52]</td>
<td>0.995</td>
<td>0.980</td>
<td>0.845</td>
<td>0.926</td>
</tr>
<tr>
<td>HIS [4]</td>
<td>-</td>
<td>0.986</td>
<td>0.881</td>
<td>0.944</td>
</tr>
<tr>
<td>Ours (ε₂)</td>
<td>0.997</td>
<td>0.987</td>
<td>0.909</td>
<td>0.951</td>
</tr>
</tbody>
</table>

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.

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

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cityscapes</th>
<th>KITTI</th>
<th>COCO</th>
<th>ADE20K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>X</td>
<td>0.027</td>
<td>0.028</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>0.048</td>
<td>0.035</td>
<td>0.081</td>
</tr>
<tr>
<td>Val</td>
<td>X</td>
<td>0.035</td>
<td>0.039</td>
<td>0.199</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>0.058</td>
<td>0.048</td>
<td>0.204</td>
</tr>
</tbody>
</table>

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
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 $u_{c}$ 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 $u_{c}$ 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 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 $i \in \{\text{person}, \text{rider}\}$ to 30 and $B_{ij}$ with $j \in \{\text{sidewalk}, \text{building}\}$ 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


[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


[18] Lei Feng, Senlin Zhu, Shuoyi 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


[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


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


