Relaxed Class-consensus Consistency for Semi-supervised Semantic Segmentation

Huayu Mai*, Rui Sun*, Feng Wu†

Deep Space Exploration Laboratory/School of Information Science and Technology,
University of Science and Technology of China
{mai556, issunrui}@mail.ustc.edu.cn, fengwu@ustc.edu.cn

Abstract

The key to semi-supervised semantic segmentation lies in how to fully exploit a large amount of unlabeled data to improve the model's generalization performance. Most methods are lured into the trap of taking each class independently (i.e., class-independent consistency) and neglecting the fact that there exist semantic dependencies among classes. In this paper, we analyze the bottlenecks of class-independent consistency inherent in previous methods and offer a fresh perspective of cooperative game theory to explicitly encourage class-consensus alignment (i.e., class-consensus consistency) between the teacher (weak augmented view) and student network (strong augmented view). We formulate classes as players in an cooperative game to model their interpretable consensus and shed light on the possibility of closer collaboration between consensus themselves and consistency regularization, yielding more comprehensive and effective supervision signals. To this end, we carefully design the class-consensus consistency without introducing any external knowledge to model class structure information which renders better interpretability, and further, prepend relaxed class-consensus consistency (RCC) to unlock the potential of modeling class consensus by relaxing the strict alignment of direct class consensus values to ranking alignment. Extensive experimental results on multiple benchmarks demonstrate that RCC performs favorably against state-of-the-art methods. Particularly in the low-data regimes, RCC achieves significant improvements.

Introduction

Semantic segmentation is a fundamental task that has achieved conspicuous achievements attributed to the recent advances in deep neural network (Long, Shelhamer, and Darrell 2015) with widespread applications such as visual understanding (Everingham et al. 2010), autonomous driving (Feng et al. 2020), etc. However, its data-driven nature makes it laborious and time-consuming to gather massive pixel-level annotations as training data. To alleviate the data-hunger issue, considerable works (Wang et al. 2023b; Yang et al. 2022a; Hu et al. 2021) have turned their attention to semi-supervised semantic segmentation. Since only limited labeled data is accessible, how to fully exploit vast unlabeled data to improve the model's generalization performance is thus extremely challenging.

Recently, the teacher-student scheme (Tarvainen and Valpola 2017) encapsulating pseudo-labeling (Lee et al. 2013; Arazo et al. 2020) and the consistency regularization paradigm (Laine and Aila 2016; Bachman, Alsharif, and Precup 2014) has dominated this field credited to their simplicity yet competitive performance. The core idea is to construct consistency regularization to align the pseudo labels generated by the teacher network under weakly augmented perturbation view with the prediction of the student network featuring strongly augmented perturbation view (where the teacher and student can be identical).

The supervision signal constituted by consistency involves hard pseudo labeling (selectively recruiting the classes with the highest confidence for training (Chen et al. 2021a)) and soft pseudo labeling (retaining scores for all classes as soft pseudo-labels to measure deviation under the presence of KL divergence (Ke et al. 2020) in Figure 1 (a)). Despite their promising results, these methods are lured into the trap of taking each class independently (i.e., *class-independent consistency*) and neglecting the fact that semantic dependencies exist among classes (i.e., class consensus). For example, the class *truck* and *car* featuring high class consensus share similar visual patterns (e.g., wheels) with each other. In this case, the i.i.d. assumption manifested in the class-independent consistency tends to be hysteresis to comprehensively probe unlabeled data. To make matters worse, the negative impact is inevitably amplified by inbuilt low-data regimes of semi-supervised semantic segmentation, compromising the capability of the model. Therefore, it is highly desirable to characterize the class consensus, that is, leverage the inter-class relationship to model structure information in pursuit of additional supervision to complement pure class-independent consistency.

In this paper, we analyze the bottlenecks of class-independent consistency inherent in previous methods and offer a fresh perspective of cooperative game theory to explicitly encourage class-consensus alignment (i.e., *class-consensus consistency*) between the teacher network (weak augmented view) and student network (strong augmented view). We formulate classes as players in a cooperative game to model their interpretable consensus and shed light on the possibility of closer collaboration between consensus them-

*Equal contribution
†Corresponding author

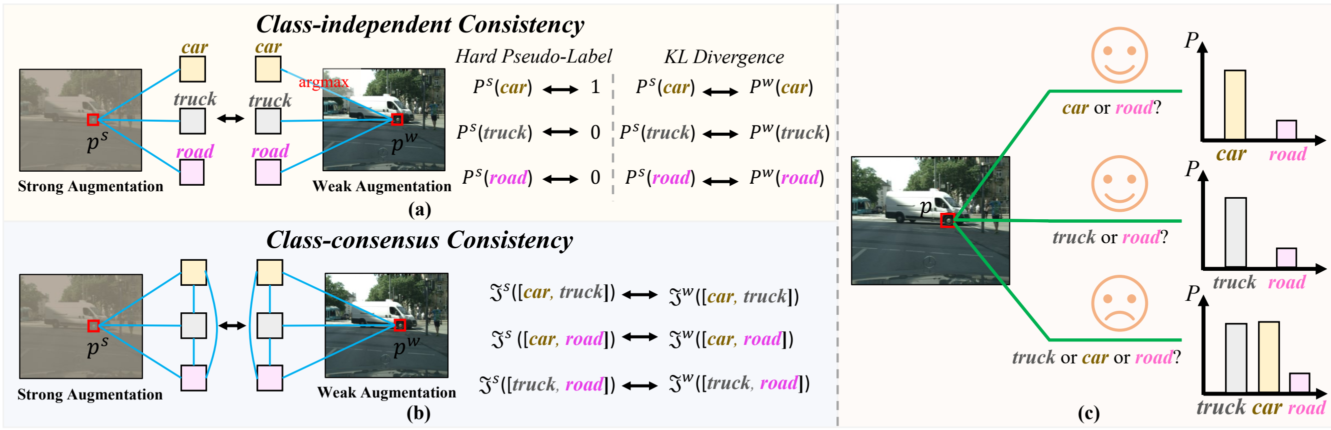

Figure 1: Illustration of our motivation. (a) The supervision signal constituted by consistency involves hard pseudo labeling and soft pseudo labeling. However, these methods are lured into the trap of taking each class independently. (b) We do not introduce any external knowledge to harness the inter-class relationship to model structure information in pursuit of additional supervision, from a fresh perspective of cooperative game theory. (c) The model inclines to be more susceptible to confusion when classes with high consensus co-occur (the coalition of classes), compared to the case when the model independently discriminates each class in the coalition.

selves and consistency regularization, yielding more comprehensive and effective supervision signals for robust semi-supervised semantic segmentation. Intuitively, if two classes share similar semantic concepts, their collaboration tends to confound the model. That is, the model inclines to be more susceptible to confusion when classes with high consensus co-occur (the coalition of classes), compared to the case when the model independently discriminates each class in the coalition. As shown in Figure 1 (c), given the presence of the class *road*, the model is capable of making the decision about the class of pixel p when the single class entity appears individually in the coalition $\{car, truck\}$, compared to discriminating the entire coalition, shedding light on the high consensus between *car* and *truck* tends to exacerbate the model's prediction uncertainty (which can be easily measured by prediction entropy). Motivated by this spirit, we construct the class coalition and develop to quantify the trend of interaction (i.e., class consensus) within a coalition via the game-theoretic interactions (i.e., Shapley interaction (Grabisch and Roubens 1999; Ren et al. 2021), which is one of the most popular theories in cooperative games) for its simplicity and efficiency. Shapley interaction measures the additional outcome brought by the coalition compared with the case in which the players work individually. In a nutshell, when a coalition has high Shapley interaction (high class consensus), it manifests a significant impact on the model's uncertainty (high prediction entropy). Thus, we can use the Shapley interaction to value possible class consensus which renders better interpretability.

To this end, we carefully design the Class-consensus Consistency (CC) without introducing any external knowledge to harness the inter-class relationship to model structure information in pursuit of additional supervision, instead of taking each class independently. In specific, we take classes as players and prediction entropy as the characteristic function in the cooperative game. Then, we use the Shapley interaction to reason about the trend of interactions inside the coalition the players constitute, aiming to learn class consensus to span the semantic space (as shown in Figure 1 (b)). It naturally comes into mind to construct supervision signal on top of class-consensus consistency in the form of L2 distance and so on. However, as a significant departure from the regular class-independent consistency which enjoys the constraints from the ground truth of labeled data as reliable support, class-consensus consistency stems from unsupervised optimization in a data-driven manner, compressing the reliability of its values. To this end, we prepend **R**elaxed **C**lass-consensus **C**onsistency (**RCC**) to strive to further unlock the potential of modeling class consensus, by relaxing the strict alignment of direct class consensus values to ranking alignment (considering that under noise interference, the relative comparability of values typically does not change, that is, the inbuilt noise-resistance capability of ranking). By constraining the ranking consistency of class consensus between the teacher and student networks, RCC enjoys special bonus from more comprehensive and effective supervision signals, enabling better description of data distribution, and achieving better exploration of unlabeled data. We find that explicitly encouraging class-consensus alignment brings remarkable improvement to already very strong pure class-independent consistency.

In this work, our contributions can be summarized as follows: (1) We analyze the bottlenecks of class-independent consistency inherent in previous methods and offer a fresh perspective on cooperative game theory to explicitly encourage class-consensus alignment. (2) We carefully design the class-consensus consistency without introducing any external knowledge to model class structure information in pursuit of additional supervision which renders better interpretability, and further, prepend relaxed class-consensus

Algorithm 1: Pseudo algorithms of RCC.

Inputs: Labeled Set $\mathcal{D}^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N^l}$, Unlabeled Set $\mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$ $(N^u \gg N^l)$
Define: Teacher Network f_T, Student Network f_S, Weak Augmentation $aug(\cdot)$, Strong Augmentation $\mathcal{A}ug(\cdot)$
Output: Student Network f_S
for each batch of $(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)$, $\boldsymbol{x}_i^u$ in $\mathcal{D}^l$, $\mathcal{D}^u$ **do**
    # Labeled Data:
    Calculate $\mathcal{L}_{sup}$ for f_S by Equation 1
        ▷ *Supervised loss*
    # Unlabeled Data:
    Obtain pseudo-labels from f_T by Equation 2
    Calculate $\mathcal{L}_{reg}$ for f_S by Equation 3
        ▷ *Class-independent Consistency*
    Calculate class consensus $\boldsymbol{J}_{ij}^w$ and $\boldsymbol{J}_{ij}^s$ for f_T and f_S respectively by Equation 7
    Calculate $\mathcal{L}_{RCC}$ for f_S by Equation 10
        ▷ *Relaxed Class-consensus Consistency*
    Gradient backward $\mathcal{L}_{sup} + \mathcal{L}_{reg} + \lambda\mathcal{L}_{RCC}$
        ▷ *Update Model*
end for

consistency (RCC) to unlock the potential of modeling class consensus by relaxing the strict alignment of direct class consensus values to ranking alignment. (3) Extensive experiments on three challenging benchmarks demonstrate that our RCC outperforms state-of-the-art semi-supervised semantic segmentation methods. Particularly in the low-data regimes, RCC achieves significant improvements.

Related Work

Semi-supervised Semantic Segmentation. The development of general semi-supervised learning and various types of semi-supervised semantic segmentation algorithms (Liang et al. 2023; Na et al. 2023; Sun et al. 2024) have brought significant benefits. These algorithms leverage the mature combination of pseudo-labeling (Chen et al. 2021a; Hu et al. 2021; Wang et al. 2022) and consistency regularization (Lai et al. 2021; Zhong et al. 2021; Ouali, Hudelot, and Tami 2020; Chen and Lian 2022) to improve performance. More recently, UniMatch (Yang et al. 2022a) acknowledges the characteristics of semantic segmentation tasks and incorporates appropriate data augmentations into FixMatch (Sohn et al. 2020), resulting in a concise yet powerful semi-supervised semantic segmentation baseline. Despite the promising results achieved by these methods, they often overlook the fact that semantic dependencies exist among different classes and tend to treat each class independently (Chen et al. 2021a; Ke et al. 2020). In this paper, we analyze the limitations of class-independent consistency found in previous methods and propose a novel perspective based on cooperative game theory to explicitly encourage class-consensus alignment.

Shapley Interaction. The Shapley interaction(Shapley 2020) was initially introduced in cooperative game theory (Chalkiadakis, Elkind, and Wooldridge 2022; Ferguson 2020) with the Shapley value. The core of the cooperative game theory is to allocate different outcomes to game individuals fairly and reasonably. The Shapley value has been theoretically demonstrated as the unique metric for impartially estimating each player's contribution in a cooperative game while ensuring the fulfillment of specific desirable axioms (Winter 2002; Marichal and Mathonet 2011). With solid theoretic foundations, Shapley interaction has recently been studied in the field of model interpretability (Datta, Sen, and Zick 2016; Lundberg and Lee 2017; Zhang et al. 2021; Ren et al. 2021; Rozemberczki et al. 2022; Djenouri et al. 2023), but there is a little exploration in the application of deep learning. In this work, we use the Shapley interaction to value possible class consensus, harnessing the inter-class relationship to model structure information in pursuit of additional supervision for robust semi-supervised semantic segmentation.

Method

In this section, we first formulate the typical teacher-student scheme of the semi-supervised semantic segmentation task and introduce the Shapley value & Shapley interaction of game theory in preliminaries. Then, we detail how the Shapley interaction can be applied to semi-supervised semantic segmentation to value potential class-class consensus. Finally, the relaxed class-consensus consistency loss is devised to harness the inter-class relationship for more effective supervision signals. In Algorithm 1, we present the pseudo algorithm of RCC to clearly summarize our method.

Preliminaries

Teacher-student Scheme. Given a labeled set $\mathcal{D}^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N^l}$ and an unlabeled set $\mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$, where $N^u \gg N^l$, semi-supervised semantic segmentation aims to train a segmentation model with limited labeled data and vast unlabeled data. The popular teacher-student scheme (Sohn et al. 2020; Yang et al. 2022a) consists of a teacher network f_T and a student network f_S. The student network is guided by two sources of supervision, including the ground truth for the labeled data and the pseudo-labels generated by the teacher network for the unlabeled data. The teacher network can either be identical to the student network or an exponentially moving average (EMA) version of it. In specific, for the labeled data, the supervised loss $\mathcal{L}_{sup}$ can be formulated as:

$$\mathcal{L}_{sup} = \frac{1}{N^l} \sum_{i=1}^{N^l} \frac{1}{HW} \sum_{j=1}^{HW} \ell_{ce}\left(\boldsymbol{y}_{ij}^l, f_S(\boldsymbol{x}_i^l)_j\right), \tag{1}$$

where H and W represent the height and width of the input image, ℓ_{ce} denotes the standard pixel-wise cross-entropy loss. For the unlabeled data, the teacher network takes the weak augmented view $aug(\boldsymbol{x}_i^u)$ as input and generates pseudo-labels $\hat{\boldsymbol{y}}_i^u$ for the student network as:

$$\hat{\boldsymbol{y}}_{ij}^u = \begin{cases} \arg\max f_T(aug(\boldsymbol{x}_i^u))_j, & c_{ij}^u > \gamma \\ \text{ignore\_index}, & \text{otherwise} \end{cases}, \tag{2}$$

where $c_{ij}^u = \max f_T(aug(\boldsymbol{x}_i^u))_j$ represents the confidence of the teacher prediction for j-th pixel in i-th input and γ denotes the confidence threshold to exclude unreliable pseudo-labels from training. As result, we can obtain the consistency regularization loss $\mathcal{L}_{reg}$ as:

$$\mathcal{L}_{reg} = \frac{1}{N^u}\sum_{i=1}^{N^u}\frac{1}{HW}\sum_{j=1}^{HW}\ell_{ce}\left(\hat{\boldsymbol{y}}_{ij}^u, f_S(\mathcal{A}ug(\boldsymbol{x}_i^u))_j\right), \quad (3)$$

where $\mathcal{A}ug(\cdot)$ means the strong augmentation. The model can learn reliable information from unlabeled data by imposing consistency regularization. The overall loss of the typical teacher-student scheme is $\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{reg}$.

Shapley Value & Shapley Interaction. The Shapley value (Shapley 2020) is a classic metric in a cooperative game. Considering a game with a set of players $\mathcal{N} = \{1, 2, ...\}$, a characteristic function $\phi(\cdot)$ is implemented to map any subset $\mathcal{T} \subseteq \mathcal{N}$ of players to a score, modeling the outcome of a game when players in $\mathcal{T}$ participate in. Formally, the Shapley value $\mathfrak{S}(u|\mathcal{N})$ measures the average marginal contribution of player u across all combinations of the players in $\mathcal{N}$:

$$\mathfrak{S}(u|\mathcal{N}) = \sum_{\mathcal{T}\subseteq\mathcal{N}/\{u\}} P(\mathcal{T})\left[\phi(\mathcal{T}\cup\{u\}) - \phi(\mathcal{T})\right], \quad (4)$$

where $P(\mathcal{T}) = \frac{1}{2^{|\mathcal{N}|-1}}$ denotes the likelihood of $\mathcal{T}$ being sampled and $\mathcal{N}/\{u\}$ denotes removing player u from $\mathcal{N}$. It is the unbiased estimation for the contribution of player u w.r.t.overall players $\mathcal{N}$.

In a cooperative game, some players tend to form a coalition and participate in the game together. The players in the coalition might well interact with each other, which brings additional contributions to the game. The Shapley interaction, derived from the Shapley value, measures the **additional contributions** brought by the coalition compared with the case in which the players work individually. Formally, for a subset of players $\mathcal{T}$, we consider $[\mathcal{T}]$ as a coalition, which is viewed as a single player in the game. The Shapley interaction for coalition $[\mathcal{T}]$ is defined as:

$$\mathfrak{I}([\mathcal{T}]) = \mathfrak{S}([\mathcal{T}]|\mathcal{N}) - \sum_{u\in\mathcal{T}}\mathfrak{S}(u|\mathcal{N}/\mathcal{T}\cup\{u\}). \quad (5)$$

Intuitively, $\mathfrak{I}([\mathcal{T}])$ reflects the trend of interactions inside players $\mathcal{T}$. The higher value of $\mathfrak{I}([\mathcal{T}])$ indicates that players in $\mathcal{T}$ interact better with each other.

Class Consensus

The typical teacher-student scheme directly imposes the consistency between the prediction of the student model and the pseudo-label generated by the teacher model, overlooking the relationship between classes. Inspired by game theory, we creatively propose to measure the inter-class relationship (i.e., class consensus) by Shapley interaction. In specific, we take all K classes as the players $\mathcal{N}$ and the prediction entropy as the characteristic function $\phi(\cdot)$. For the prediction $\boldsymbol{p} \in \mathbb{R}^K$ of a pixel, the entropy w.r.t.a subset of classes $\mathcal{T}$ is defined as:

$$\phi(\mathcal{T}) = -\sum_{u\in\mathcal{T}}\widetilde{\boldsymbol{p}}_u\log\widetilde{\boldsymbol{p}}_u, \quad \widetilde{\boldsymbol{p}}_u = \frac{\exp(\boldsymbol{p}_u)}{\sum_{u\in\mathcal{T}}\exp(\boldsymbol{p}_u)}. \quad (6)$$

From Equation 4 and Equation 5, the Shapley interaction of classes u and v (i.e., class consensus between u and v) is:

$$\begin{aligned}\mathfrak{I}([u,v]) &= \mathfrak{S}([u,v]|\mathcal{N}) - \mathfrak{S}(u|\mathcal{N}/v) - \mathfrak{S}(v|\mathcal{N}/u) \\ &= \sum_{\mathcal{T}\subseteq\mathcal{N}/\{u,v\}}\frac{1}{2^{K-2}}[\phi(\mathcal{T}\cup\{[u,v]\}) + \phi(\mathcal{T}) \\ &\quad - \phi(\mathcal{T}\cup\{u\}) - \phi(\mathcal{T}\cup\{v\})].\end{aligned} \quad (7)$$

For example, suppose we have three classes in total: *car*, *truck* and *road*. Given the pixel of the wheel, the model will confidently determine whether it belongs to the *car* (*truck*) or *road* category, i.e., the prediction entropy $\phi(\{car, road\})$ ($\phi(\{truck, road\})$) is small. However, when the model needs to make a three-class classification among *car*, *truck* and *road*, it becomes uncertain, i.e., the prediction entropy $\phi(\{car, truck, road\})$ is large. According to Equation 7, $\mathfrak{I}([car, truck])$ will be a relatively large value, which is in line with the reality that *car* and *truck* share high semantic consensus. Intuitively, the co-occurrence of two classes with a high semantic consensus will confuse the model, leading to a large Shapley interaction. This is why we use Shapley interaction to measure the class consensus.

Relaxed Class-consensus Consistency

After defining the class consensus, we can obtain a set of $\boldsymbol{J}_{ij} = \{\mathfrak{I}([u,v])\}$ ($|\boldsymbol{J}_{ij}| = C_2^K$) for each pixel $\boldsymbol{x}_{ij}$. Straightforwardly, we can impose the consistency between the class-consensus $\boldsymbol{J}_{ij}^w$ of the teacher network and the $\boldsymbol{J}_{ij}^s$ of the student network resorting to L2 distance as:

$$\mathcal{L}_{CC} = \frac{1}{N^u}\sum_{i=1}^{N^u}\frac{1}{HW}\sum_{j=1}^{HW}\ell_{L2}\left(\boldsymbol{J}_{ij}^w, \boldsymbol{J}_{ij}^s\right). \quad (8)$$

However, unlike regular class-independent consistency, which enjoys the constraints from the ground truth of labeled data as reliable support, class-consensus consistency stems from unsupervised optimization in a data-driven manner, making the strict alignment full of noise. To unlock the potential of modeling class consensus, we propose the relaxed class-consensus consistency by relaxing the strict alignment of direct class consensus values to ranking alignment:

$$\ell_{rk}\left(\boldsymbol{J}^w, \boldsymbol{J}^s\right) = \sum_{k}^{C_2^K}\left[1 - \frac{\min(\mathcal{R}(k,\boldsymbol{J}^w), \mathcal{R}(k,\boldsymbol{J}^s))}{\max(\mathcal{R}(k,\boldsymbol{J}^w), \mathcal{R}(k,\boldsymbol{J}^s))}\right], \quad (9)$$

$$\mathcal{L}_{RCC} = \frac{1}{N^u}\sum_{i=1}^{N^u}\frac{1}{HW}\sum_{j=1}^{HW}\ell_{rk}\left(\boldsymbol{J}_{ij}^w, \boldsymbol{J}_{ij}^s\right), \quad (10)$$

where ℓ_{rk} is a ranking order-based loss function and $\mathcal{R}(k, \boldsymbol{J})$ refers to the ranking of the instance k in the class consensus list $\boldsymbol{J}$. Obviously, the ranking function $\mathcal{R}(k, \boldsymbol{J})$ is non-differentiable, resulting in the inability to be optimized with gradient-based approaches. Rewrite $\mathcal{R}(k, \boldsymbol{J})$ as:

$$\mathcal{R}(k,\boldsymbol{J}) = 1 + \sum_{l\neq k}\mathbb{I}\{\boldsymbol{J}_l - \boldsymbol{J}_k\}, \quad (11)$$

Method	ResNet-50					ResNet-101				
	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)	1/16(92)	1/8(183)	1/4(366)	1/2(732)	Full(1464)
Sup.-only	44.0	52.3	61.7	66.7	72.9	45.1	55.3	64.8	69.7	73.5
FixMatch	60.1	67.3	71.4	73.7	76.9	63.9	73.0	75.5	77.8	79.2
iMAS	–	–	–	–	–	68.8	75.3	79.1	80.2	82.0
AugSeg	64.2	72.1	76.1	77.4	78.8	71.0	75.4	78.8	80.3	81.3
DGCL	–	–	–	–	–	70.4	77.1	78.7	79.2	81.5
CSS	68.0	71.9	74.9	77.6	–	–	–	–	–	–
LOGICDIAG	–	–	–	–	–	73.2	76.6	77.9	79.3	–
NP-SemiSeg	65.7	72.3	75.7	77.4	–	–	–	–	–	–
DAW	68.5	73.1	76.3	78.6	79.7	74.8	77.4	79.5	80.6	81.5
Switch	70.7	74.5	76.4	77.6	78.1	–	–	–	–	–
UniMatch	67.4	71.9	75.3	78.0	79.3	73.5	75.4	78.7	80.2	81.9
RCC (Ours)	**71.8**	**74.7**	**76.9**	**78.9**	**80.3**	**75.3**	**77.9**	**79.8**	**81.0**	**82.1**
$\Delta \uparrow$	**+27.8**	**+22.4**	**+15.2**	**+12.2**	**+7.4**	**+30.2**	**+22.6**	**+15.0**	**+11.3**	**+8.6**

Table 1: Quantitative results of different SSL methods on Pascal *classic* set. We report mIoU (%) under various partition protocols and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Method	ResNet-50			ResNet-101		
	1/16(662)	1/8(1323)	1/4(2646)	1/16(662)	1/8(1323)	1/4(2646)
Sup.-only	62.4	68.2	72.3	67.5	71.1	74.2
FixMatch	70.6	73.9	75.1	74.3	76.3	76.9
ST++	72.6	74.4	75.4	74.5	76.3	76.6
U^2PL	–	–	–	77.2	79.0	79.3
AugSeg	74.6	75.9	77.1	77.0	77.3	78.8
iMAS	75.9	76.7	77.1	77.2	78.4	79.3
CFCG	75.0	77.1	77.7	76.8	79.1	79.9
NP-SemiSeg	73.4	76.5	76.7	–	–	–
DAW	76.2	77.6	77.4	78.5	78.9	79.6
UniMatch	75.8	76.9	76.8	78.1	78.4	79.2
RCC (Ours)	**76.9**	**78.1**	**78.1**	**79.1**	**79.4**	**80.2**
$\Delta \uparrow$	**+14.5**	**+9.9**	**+5.8**	**+11.6**	**+8.3**	**+6.0**

Table 2: Quantitative results of different SSL methods on Pascal *blender* set. We report mIoU (%) under various partition protocols and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

where $\mathbb{I}(\cdot)$ serves as an Indicator function which equals to 1 when $\boldsymbol{J}_l - \boldsymbol{J}_k > 0$. To make $\mathcal{L}_{RCC}$ differentiable, we utilize a sigmoid function in place of the Indicator function.

Finally, the overall learning objective of our RCC is derived as:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{reg} + \lambda \mathcal{L}_{RCC}, \tag{12}$$

where the λ is the trade-off weight.

Experiments

Experimental Setup

Datasets. (1) **PASCAL VOC 2012** (Everingham et al. 2010) is an object-centric semantic segmentation dataset, containing 20 object classes in the foreground and a background class with 1,464 and 1,449 finely annotated images for training and validation, respectively. Many researches (Chen et al. 2021b; Hu et al. 2021) augment the original training set (i.e., *classic*) with additional 9,118 coarsely annotated images in SBD (Hariharan et al. 2011) to get a *blender* training set. (2) **Cityscapes** (Cordts et al. 2016) is an urban scene understanding dataset consisting of 2,975 images for training and 500 images for validation. The initial 30 semantic classes are re-mapped into 19 classes for the semantic segmentation task.

Implementation Details. For a fair comparison, we use ResNet-50/101 (He et al. 2016) pretrained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) as the backbone and DeepLabv3+ (Chen et al. 2018) as the decoder. The crop size is set as 513×513 for PASCAL and 801×801 for Cityscapes, respectively. We adopt stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001 for PASCAL and 0.005 for Cityscapes. Polynomial Decay learning rate policy is applied throughout the whole training. The strong augmentation $\mathcal{A}ug(\cdot)$ contains random color jitter, grayscale and Gaussian blur. The weak augmentation $aug(\cdot)$ consists of random crop, resize and horizontal flip. For computational efficiency, we focus on the consensus between the top-4 classes for each pixel, based on our observation that in every prediction, the top-4 classes have occupied almost all weight. We set the trade-off weight $\lambda = 0.2$ for all experiments. The model is trained for 80 epochs on PASCAL and 240 epochs on Cityscapes with a batch size of 8, using $8\times$ RTX 3090 GPUs.

Comparison with State-of-the-art Methods

We utilize the widely adopted consistency regularization framework UniMatch (Yang et al. 2022a) as our baseline, where both the teacher and student networks are identical. We conduct evaluations of our method across the PASCAL datasets (*classic* and *blender*) as well as the Cityscapes dataset, employing both ResNet-50 and ResNet-101 backbones across various partition protocols. Our evaluations involve exhaustive comparisons with SOTA methods (Sohn et al. 2020; Zhao et al. 2023a,b; Wang et al. 2023d,a; Liang et al. 2023; Wang et al. 2023c; Na et al. 2023; Yang et al. 2022b; Wang et al. 2022; Li et al. 2023a,b; Ma et al. 2023). The consistently dominant performance under all partition protocols with different backbones on all datasets proves the effectiveness of RCC.

Results on PASCAL. Table 1 and Table 2 illustrate the comparison between our method and the SOTA methods on both the PASCAL *classic* and *blender* datasets. In contrast to the supervised-only (*Sup.-only*) model, our approach demonstrates notable performance enhancements, indicating effective utilization of information from unlabeled data. Furthermore, consistent and significant performance improvements are observed compared to the baseline method, UniMatch. Specifically, employing our approach yields performance scores of 71.8% and 75.3% under $1/16(92)$ partition on the *classic* set with ResNet-50 and ResNet-101 backbones, respectively, resulting in a boost over the baseline by 4.4% and 1.8%. These findings underscore the potent information extraction capabilities of our RCC, especially in scenarios with extremely limited labeled data.

Results on Cityscapes. Table 3 tabulates a comparative analysis of RCC against state-of-the-art (SOTA) methods on the Cityscapes dataset. Notably, utilizing the ResNet-50 backbone, RCC demonstrates significant performance improvements over the *Sup.-only* model by 12.4%, 7.7%, 6.2%, and 3.2% under $1/16$, $1/8$, $1/4$, and $1/2$ partition protocols, respectively. Moreover, compared to the recent and competitive contrastive method Co-Train (Li et al. 2023b), our approach maintains superiority in performance. For instance, with the ResNet-101 backbone under the $1/16$ partition protocol, our method exhibits a 2.3% performance improvement, underscoring its superiority over contrastive learning techniques.

Qualitative Results. We assess the qualitative results of our method against various state-of-the-art competitors on the PASCAL dataset. As shown in Figure 2, RCC displays notably enhanced segmentation performance, particularly evident in capturing fine-grained details, such as the depiction of objects like clustered dogs and the man on horseback.

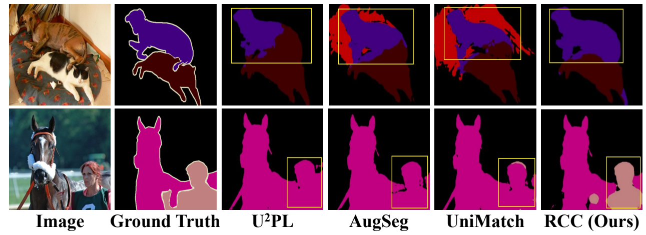

Figure 2: Qualitative comparison with different methods. Significant improvements are marked with yellow boxes.

Leveraging the relaxed class-consensus consistency, RCC demonstrates superior capabilities across most scenarios.

Ablation Study and Analysis

To look deeper into our method, we perform a series of ablation studies on PASCAL *classic* set under $1/16$ (92) partition protocol with ResNet-50 to analyze our RCC. Note that our baseline is UniMatch (Yang et al. 2022a), which relies on pure class-independent consistency.

Effectiveness of Components. As tabulated in Table 4, we report the performance under $1/16$ (92) and Full (1464) partition protocol to justify the correctness and effectiveness of each component. Note that *CC* represents class-consensus consistency equipped with supervision signal constructed by L2 distance, and *Relaxed* serves to relax the strict alignment (e.g., L2 distance) of *CC* to ranking alignment. In other words, *CC* and *Relaxed* constitute our final model RCC (relaxed class-consensus consistency). (1) From the comparison between the 1^{st} row and 2^{nd} row of Table 4, we find that the introduction of class consensus to model class-consensus consistency (L2 loss in default) in pursuit of additional supervision achieves clear performance gains (e.g., 2.8% in mIoU under $1/16$ (92) partition protocol). We conclude that the performance gain comes from explicitly encouraging class-consensus alignment (class-consensus consistency) between the teacher and student network equipped with the cooperative game theory. (2) The addition of relaxing the strict alignment of direct class consensus values to ranking alignment (2^{nd} vs.3^{rd} row) also contributes to a remarkable performance gain compared with in Table 4 (e.g., 1.6% in mIoU under $1/16$ (92) partition protocol). The improvements can be mainly ascribed our RCC enjoys the inbuilt noise-resistance capability to further unlock the potential of modeling class consensus, enabling a better description of data distribution.

Analysis of Constructed Loss. To explore the effectiveness of different strategies to construct supervision signal (i.e., loss) on top of class-consensus consistency (*CC*), we conduct experiments in Table 5. Among them, L2, CE, and KL belong to the strict alignment of direct class consensus values based on class-consensus consistency. However, as a significant departure from the regular class-independent consistency which enjoys the constraints from the ground truth of labeled data as reliable support, class-consensus consistency stems from unsupervised optimization in a data-driven manner, compressing the reliability of its values. Therefore, we

Method	ResNet-50				ResNet-101			
	1/16(186)	1/8(372)	1/4(744)	1/2(1488)	1/16(186)	1/8(372)	1/4(744)	1/2(1488)
Sup.-only	63.3	70.2	73.1	76.6	66.3	72.8	75.0	78.0
FixMatch	72.6	75.7	76.8	78.2	74.2	76.2	77.2	78.4
AEL	74.0	75.8	76.2	–	75.8	77.9	79.0	80.3
AugSeg	73.7	76.4	78.7	79.3	75.2	77.8	79.5	80.4
iMAS	74.3	77.4	78.1	79.3	–	–	–	–
ESL	–	–	–	–	75.1	77.1	78.9	80.4
Co-Train	–	76.3	77.1	–	75.0	77.3	78.7	–
NP-SemiSeg	73.0	77.1	78.8	78.7	–	–	–	–
Switch	–	–	–	–	76.8	78.4	79.4	80.5
DAW	75.2	77.5	79.1	79.5	76.6	78.4	79.8	80.6
UniMatch	75.0	76.8	77.5	78.6	76.6	77.9	79.2	79.5
RCC (Ours)	**75.7**	**77.9**	**79.3**	**79.8**	**77.3**	**78.7**	**80.3**	**80.9**
$\Delta \uparrow$	**+12.4**	**+7.7**	**+6.2**	**+3.2**	**+11.0**	**+5.9**	**+5.3**	**+2.9**

Table 3: Quantitative results of different SSL methods on Cityscapes. We report mIoU (%) under various partition protocols and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Baseline	*CC*	*Relaxed*	mIoU(92)	mIoU(1464)
✓			67.4	79.3
✓	✓		70.2	79.9
✓	✓	✓	**71.8**	**80.3**

Table 4: Ablation studies of different components on PASCAL *classic*.

CC	mIoU
L2	70.2
CE	69.9
KL	70.0
Relaxed	**71.8**

Table 5: Ablation of losses on CC.

Method	mIoU	FPS
Baseline	67.4	9.87
U^2PL	69.5	7.18
RCC (Ours)	**71.8**	8.23

Table 6: Efficiency of RCC and other competitors.

prepend the Relaxed strategy to relax the strict alignment to ranking alignment (considering that under noise interference, the relative comparability of values typically does not change, that is, the inbuilt noise-resistance capability of ranking), achieving a distinctively best performance.

Analysis of Efficiency. Table 6 summarizes the training efficiency of our RCC compared to other competing methods. U^2PL (Wang et al. 2022), a classic and competitive method based on contrastive learning, shares the same design philosophy as ours in seeking more supervision signals to explore unlabeled data. Note that we report the reproduced results for U^2PL based on our baseline (i.e., UniMatch). We can vividly observe that our RCC achieves a clear lead over U^2PL, both in terms of performance and efficiency. Compared to the baseline, RCC has an acceptable decrease in training efficiency, outweighed by the substantial improvement achieved attributed to the modeling of class consensus.

Hyperparameter Evaluations. (1) In fact, if we consider the consensus between all K classes, the complexity is unacceptable. Therefore, we calculate the top-N class consensus in each prediction. In the implementation, we take top-N classes in predictions of the teacher network and retrieve predictions of the student network for the corresponding classes. We take ablation experiments on N (Table 7), showing the best trade-off between performance and cost is achieved with $N = 4$. (2) λ controls the relative importance

N	mIoU	FPS	FLOPs (G)
3	70.3	9.21	917.3
4	71.8	8.23	928.2
5	72.0	7.36	959.7
6	71.9	6.51	1051.5

Table 7: Ablation of Top-N.

λ	mIoU
0.1	70.5
0.2	**71.8**
0.5	71.5
1.0	70.0

Table 8: Ablation of λ.

of the relaxed class-consensus consistency loss; our model achieves much better performance when $\lambda = 0.2$ as shown in Table 8.

Conclusion

In this paper, we analyze the bottlenecks of class-independent consistency and offer a fresh perspective on cooperative game theory to explicitly encourage class-consensus alignment. We carefully design the class-consensus consistency to model class structure information in pursuit of additional supervision, and further prepend relaxed class-consensus consistency (RCC) to relax the strict alignment of direct class consensus values to ranking alignment. Extensive experimental results on challenging benchmarks show the effectiveness our proposed RCC.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27.

Chalkiadakis, G.; Elkind, E.; and Wooldridge, M. 2022. *Computational aspects of cooperative game theory*. Springer Nature.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021a. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021b. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Chen, Z.; and Lian, Z. 2022. Semi-supervised Semantic Segmentation via Prototypical Contrastive Learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6696–6705.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, 598–617. IEEE.

Djenouri, Y.; Belbachir, A. N.; Michalak, T.; and Yazidi, A. 2023. Shapley Deep Learning: A Consensus for General-Purpose Vision Systems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1224–1233.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; and Dietmayer, K. 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3): 1341–1360.

Ferguson, T. S. 2020. *A course in game theory*. World Scientific.

Grabisch, M.; and Roubens, M. 1999. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28: 547–565.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, 991–998. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, H.; Wei, F.; Hu, H.; Ye, Q.; Cui, J.; and Wang, L. 2021. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34: 22106–22118.

Ke, Z.; Qiu, D.; Li, K.; Yan, Q.; and Lau, R. W. 2020. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, 429–445. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lai, X.; Tian, Z.; Jiang, L.; Liu, S.; Zhao, H.; Wang, L.; and Jia, J. 2021. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1205–1214.

Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896.

Li, S.; He, Y.; Zhang, W.; Zhang, W.; Tan, X.; Han, J.; Ding, E.; and Wang, J. 2023a. CFCG: Semi-Supervised Semantic Segmentation via Cross-Fusion and Contour Guidance Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16348–16358.

Li, Y.; Wang, X.; Yang, L.; Feng, L.; Zhang, W.; and Gao, Y. 2023b. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*.

Liang, C.; Wang, W.; Miao, J.; and Yang, Y. 2023. Logic-induced Diagnostic Reasoning for Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16197–16208.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ma, J.; Wang, C.; Liu, Y.; Lin, L.; and Li, G. 2023. Enhanced Soft Label for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1185–1195.

Marichal, J.-L.; and Mathonet, P. 2011. Weighted Banzhaf power and interaction indexes through weighted approximations of games. *European journal of operational research*, 211(2): 352–358.

Na, J.; Ha, J.-W.; Chang, H. J.; Han, D.; and Hwang, W. 2023. Switching Temporary Teachers for Semi-Supervised Semantic Segmentation. *arXiv preprint arXiv:2310.18640*.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

Ren, J.; Zhang, D.; Wang, Y.; Chen, L.; Zhou, Z.; Chen, Y.; Cheng, X.; Wang, X.; Zhou, M.; Shi, J.; et al. 2021. A unified game-theoretic interpretation of adversarial robustness. *arXiv preprint arXiv:2103.07364*.

Rozemberczki, B.; Watson, L.; Bayer, P.; Yang, H.-T.; Kiss, O.; Nilsson, S.; and Sarkar, R. 2022. The Shapley Value in Machine Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 5572–5579. International Joint Conferences on Artificial Intelligence Organization.

Shapley, L. 2020. 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307–317. *Classics in Game Theory; Princeton University Press: Princeton, NJ, USA*, 69–79.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Sun, B.; Yang, Y.; Zhang, L.; Cheng, M.-M.; and Hou, Q. 2024. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3097–3107.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Wang, C.; Xie, H.; Yuan, Y.; Fu, C.; and Yue, X. 2023a. Space Engage: Collaborative Space Supervision for Contrastive-based Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 931–942.

Wang, J.; Massiceti, D.; Hu, X.; Pavlovic, V.; and Lukasiewicz, T. 2023b. NP-SemiSeg: when neural processes meet semi-supervised semantic segmentation. In *International Conference on Machine Learning*, 36138–36156. PMLR.

Wang, J.; Massiceti, D.; Hu, X.; Pavlovic, V.; and Lukasiewicz, T. 2023c. NP-SemiSeg: when neural processes meet semi-supervised semantic segmentation. In *International Conference on Machine Learning*, 36138–36156. PMLR.

Wang, X.; Zhang, B.; Yu, L.; and Xiao, J. 2023d. Hunting Sparsity: Density-Guided Contrastive Learning for Semi-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3114–3123.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4248–4257.

Winter, E. 2002. The shapley value. *Handbook of game theory with economic applications*, 3: 2025–2054.

Yang, L.; Qi, L.; Feng, L.; Zhang, W.; and Shi, Y. 2022a. Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation. *arXiv preprint arXiv:2208.09910*.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022b. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4268–4277.

Zhang, H.; Xie, Y.; Zheng, L.; Zhang, D.; and Zhang, Q. 2021. Interpreting multivariate shapley interactions in dnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10877–10886.

Zhao, Z.; Long, S.; Pi, J.; Wang, J.; and Zhou, L. 2023a. Instance-specific and Model-adaptive Supervision for Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23705–23714.

Zhao, Z.; Yang, L.; Long, S.; Pi, J.; Zhou, L.; and Wang, J. 2023b. Augmentation Matters: A Simple-yet-Effective Approach to Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11350–11359.

Zhong, Y.; Yuan, B.; Wu, H.; Yuan, Z.; Peng, J.; and Wang, Y.-X. 2021. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7273–7282.