Conservative-Progressive Collaborative Learning for Semi-Supervised Semantic Segmentation

Siqi Fan, Fenghua Zhu, Senior Member, IEEE, Zunlei Feng, Yisheng Lv, Senior Member, IEEE, Mingli Song, Senior Member, IEEE, and Fei-Yue Wang, Member, IEEE

Abstract—Pseudo supervision is regarded as the core idea in semi-supervised learning for semantic segmentation, and there is always a trade-off between utilizing only the high-quality pseudo labels and leveraging all the pseudo labels. Addressing that, we propose a novel learning approach, called Conservative-Progressive Collaborative Learning (CPCL), among which two predictive networks are trained in parallel, and the pseudo supervision is implemented based on both the agreement and disagreement of the two predictions. One network seeks common ground via intersection supervision and is supervised by the high-quality labels to ensure a more reliable supervision, while the other network reserves differences via union supervision and is supervised by all the pseudo labels to keep exploring with curiosity. Thus, the collaboration of conservative evolution and progressive exploration can be achieved. To reduce the influences of the suspicious pseudo labels, the loss is dynamic re-weighted according to the prediction confidence. Extensive experiments demonstrate that CPCL achieves state-of-the-art performance for semi-supervised semantic segmentation.

Index Terms—Semantic segmentation, semi-supervised learning, pseudo supervision.

I. INTRODUCTION

IMAGE semantic segmentation is a core computer vision task, and plays an important role for the environment perceptions of many intelligent systems [1]. With the wide adoption of deep supervision learning, current state-of-the-art methods have performed satisfactory predictions given sufficient pixel-level labeled data. However, the performance would drop significantly when the labeled data is limited. It is well-known that the acquisition of the large amount of pixel-level labeled data for segmentation is costly and time consuming. For example, the annotation process costs more than 1.5h for a human annotator to label a high-resolution image of urban street scenes [2], which is 15 times and 60 times larger than that of region-level and image-level labels [3]. Therefore, this paper study semi-supervised learning (SSL) for semantic segmentation which leverages the large amount of unlabeled data to alleviate the labeled data dependency.

Though SSL is still in its initial stage, some pioneering works have been accomplished. Entropy minimization and consistency regularization are two typical approaches which are proposed in recent years. Entropy minimization methods can generate the pseudo labels on unlabeled images using the model trained on the labeled images, and retrain the model with the expand training data [4], [5]. Consistency regularization methods aim to enforce an invariance of the predictions with various perturbations [6], [7], [8], including the perturbation on input images [9], features [10] and network [11]. During the process of pseudo supervision, both the two lines of approaches are plagued with confirmation...
bias [12], which are caused by the errors in pseudo labels. Addressing that, prediction confidence is adopted by a few recent works as indicator to select reliable pseudo labels [5], [13]. Though setting the threshold of the confidence can avoid performance degradation, a large portion of unlabeled data is wasted. For example, only 27%-36% pixels can be used in the experiments of Cityscapes [5], which is far from the goal of SSL. There is always a tradeoff between quality and quantity. Besides prediction confidence, the agreement of multiple predictions can also be used as an indicator, which is the basic idea of some mutual learning methods [11], [14], [15].

Teacher-Student and Student-Student are two typical mutual learning manners, and they are both faced with challenges. The inequality in Teacher-Student may cause misleading and limitation, while the two networks trained in Student-Student manner may get into model coupling problem that reduces the mutual learning to self-training and hinders the further improvement [16].

Considering these limitations, we propose the Conservative-Progressive Collaborative Learning (CPCL), which trains two paralleled networks via the intersection supervision using only high-quality pseudo labels and union supervision using large-quantity pseudo labels, respectively, and achieves the collaboration of conservative evolution and progressive exploration, as shown in Fig. 1. The two networks share the same structure, but are initialized differently. The pseudo labels are generated according to the predictions of the two networks on the unlabeled images. Specifically, one network is conservative and supervised by the more reliable ‘intersection’ parts, while the other network is progressive and is under the pseudo supervision of the ‘union’ parts to explore the disagreement predictions. Therefore, the two networks are trained under the heterogeneous knowledge, and the model coupling problem can be reduced to a certain extent. In addition, the prediction confidence is used for loss re-weighting to deal with the ineluctable noisy pseudo labels.

The main contributions of our work are listed as follows:

- We propose the Conservative-Progressive Collaborative Learning (CPCL), which seeks common ground via the intersection supervision while reserving differences through the union supervision, and achieves the collaboration of conservative evolution and progressive exploration. The intersection supervision utilizes the high-quality pseudo labels, and the union supervision leverages the large quantity of pseudo labels. The two paralleled networks are trained using heterogeneous knowledge to reduce the model coupling problem.

- The class-wise disagreement indicator and the confidence-based dynamic loss are proposed. The indicator is proposed and calculated to determine the pseudo labels of the disagreement parts. The confidence-based dynamic loss utilizes prediction confidence for loss re-weighting, instead of thresholding.

- Extensive experimental studies are conducted with various settings on two widely used semantic segmentation datasets, Cityscapes and PASCAL VOC 2012. CPCL shows state-of-the-art performance for semi-supervised segmentation on two benchmarks, and the effectiveness of our approach is verified.

II. RELATED WORK

In this paper, we focus on how to leverage the available unlabeled images to derive additional training signals. From the perspective of architecture, the comparison of the single-branch self-training, two typical double-branch mutual learning approaches and the proposed conservative-progressive collaborative learning is shown in Fig. 2. Different from others, the two branches are equally trained using heterogeneous knowledge to reduce the misleading and model coupling problem. From the viewpoint of the prior knowledge, SSL methods are mainly based on three types knowledge, including entropy minimization, consistency regularization and agreement-based knowledge.

A. Entropy Minimization

Entropy minimization encourages the model to make low-entropy predictions on unlabeled data. The predictive entropy for a model is defined as

$$H = \sum_{c=1}^{C} p_c \log p_c,$$

where $C$ is the number of classes and $p_c$ is the probability for class $c$. The entropy would decrease when the model makes confident predictions on unlabeled data. Self-training based on pseudo labeling is a typical technique for entropy minimization, which utilizes the pseudo labels for fine-tuning the segmentation model. The pseudo supervision of the online self-training methods is propagated backward immediately [17], while the relabeling-retrain process of the offline self-training methods can be iterated several times [5], [15], [18], [19], [20], [21], [22]. The quality of the pseudo labels is the key of entropy minimization, and most of methods select the predictions with higher confidence values as the
pseudo labels. Earlier works regard the predicted probability as the prediction confidence [23], [24], [25], and some of them set hard threshold for selection [17], [18]. The pseudo label $l$ is defined as

$$l = \begin{cases} \tau & f(\tau|x) > T \\ \text{ignored} & \text{otherwise} \end{cases}$$  \tag{2}

where $\tau = \arg \max_x f(c|x)$, and $T$ is the threshold. The GAN-based methods utilize the discriminator as the selector to approximate the confidence, and select high-confident predictions by distinguishing the ground-truth and predictions [5], [26], [27]. Other works regard averaged predictions as pseudo labels based on the assumption that the averaged predictions are more confident. FastSWA [28] averages models between epochs, while Mean-Teacher [6] turns to exponential moving average. Temporal Model [29] leverages the predictions over epochs. These methods assume that the labeled data is enough to let the trained model produce the pseudo labels, which is invalid in low-data regime, and the pseudo labels with noise might result in self-error accumulation via training.

B. Consistency Regularization

Consistency regularization enforces the model remain unchanged when various perturbations occur. Input perturbations [30] generate the inconsistency by altering the inputs, which is similar to data augmentation. Data interpolation is adopted by MixMatch [31] and ReMixMatch [32]. CutMix [9] extends Cutout [33] and Mixup [34] by replacing the image regions with patches from another image, and is used in several methods [16], [35]. VAT [7] generates adversarial perturbations that alter the predictions most. FixMatch [36] and PseudoSeg [37] encourage the consistency of the predictions under strongly-augmented and the weakly-augmented. Feature perturbations are adopted by CCT [10], which alters the encoder’s outputs and enforces the consistency between the predictions of multiple auxiliary decoders. Other works either leverage the Dropout to perturb the predictions [29], [38], or perform network perturbations using two same and independently-initialized network branches [35], [39]. The performance of these methods heavily depends on the perturbations they used.

C. Mutual Learning Based on Agreement and Disagreement

Mutual learning trains two networks in parallel, which enables the utilization of multiple predictions to generate the pseudo supervision. Dual Student [11] trains two networks online, and one teaches another if it has more certain prediction. The two networks in Deep Co-training [14] are trained with the objective to maximize the agreement between them. DMT [15] updates one network with the fixed pseudo labels from another iteratively, and dynamically re-weights the loss depending on the degree of the disagreement. To deal with the coupling issue, Robust Mutual Learning [16] generates pseudo labels from the mean-teachers, which transfers the supervision in an indirect manner, and introduces the self-rectification for the labels according to the internal knowledge. In addition, the cross-image label agreement and disagreement are leveraged to mine the cross-image semantic relations for weakly supervised semantic segmentation [40].

Different from them, the two networks in CPCL are trained under the pseudo supervision of the ‘intersection’ and ‘union’ parts, achieving the collaboration of the conservative evolution and the progressive exploration, and the loss is re-weighted according to the prediction confidence to reduce the influences of the suspicious pseudo labels.

III. METHODOLOGY

In this section, we firstly expound the problem definition, and then propose the conservative-progressive collaborative learning framework. To generate the pseudo labels for the progressive union supervision, a label determination approach based on class-wise disagreement indicator for disagreement parts is proposed. We further present the confidence-based dynamic loss on the basis of the prediction confidence to deal with the unreliable pseudo labels.

A. Problem Definition

The data provided in the SSL tasks including labeled data and unlabeled data, and the unlabeled training data set is much larger than the labeled one. Without loss of generality, let $D_l = \{(X_l^1, G^1), \ldots, (X_l^n, G^n), \ldots, (X_l^N, G^N)\}$ denote the $N$ labeled data, while the $M$ unlabeled data is represented as $D_u = \{X_u^1, \ldots, X_u^m, \ldots, X_u^M\}$. $X_u^m$ and $G^u$ are the labeled input image and the corresponding label for the semantic segmentation, and $X_u^m$ is the $m$-th unlabeled input image. Specifically, $X_u^m, G^u \in \mathbb{R}^{H \times W}$, and $G^u \in \mathbb{R}^{C \times H \times W}$.

To perform well on the test data, the key problem is to leverage the large number of unlabeled data $D_u$ for the training of the segmentation network $f_\theta(X)$. In this work, the main architecture contains two segmentation networks, $f_\theta(X)$ and $f_\theta_p(X)$. While utilizing the labeled data $D_l$, the networks are trained in a traditional supervised manner, the proposed CPCL is introduced for the training on the unlabeled data $D_u$.

B. Conservative-Progressive Collaborative Learning

To not only take the advantage of the high-quality labels but also make the full use of the large quantity of pseudo labels, we propose a novel semi-supervised learning approach for semantic segmentation, named conservative-progressive collaborative learning. It is composed of two paralleled branches, one is for conservative evolution, and the other is for progressive exploration. The conservative branch is under the intersection pseudo supervision to ensure a more reliable supervision by utilizing high-quality labels, while the progressive branch is supervised by the union pseudo labels to achieve the exploration of the disagreement parts. The framework of the proposed CPCL is shown in Fig. 3.

The segmentation network in each branch shares the same structure and has different initialization. Two unlabeled images, $X_1$ and $X_2$, are input, and an augmented image $X_s$ is generated utilizing the strong augmentation methods, e.g., CutMix, CutOut and CowMix [9]. CutMix is adopted in this
The framework of the proposed conservative-progressive collaborative learning (CPCL). Two networks are trained in parallel, which share the same structure but are initialized differently. The pseudo supervision is generated on the basis of the agreement and disagreement instead of utilizing the output of each branch directly. The conventional branch and the progressive branch are supervised by the union pseudo supervision utilizing the large quantity of pseudo labels for exploration. The two branches seek common ground while reserving differences, and achieve the collaboration of the conservative and the progressive. The visualization results in the figure are output from the networks at the later training stage.

Paper, which requires two input images and selects the output pixel based on a random mask,

$$m_{ix}(a, b, \text{mask}) = (1 - \text{mask}) \odot a + \text{mask} \odot b,$$

where $a$ and $b$ are the two inputs. The augmented image $X_s$ is generated as

$$X_s = m_{ix}(X_1, X_2, \text{mask}).$$

Take the conservative branch as example, the three images are fed into the network $f_{bc}(X)$, and the segmentation predictions are $Y^1_{cw}$, $Y^2_{cw}$, and $Y_{cs}$.

$$Y^1_{cw} \leftarrow \arg \max_y f_{bc}(y|X_1),$$

$$Y^2_{cw} \leftarrow \arg \max_y f_{bc}(y|X_2),$$

$$Y_{cs} \leftarrow \arg \max_y f_{bc}(y|X_s).$$

The output $Y^1_{cw}$ and $Y^2_{cw}$ are further mixed using the same mix mask of the input, and the result is $Y_{cw}$,

$$Y_{cw} = m_{ix}(Y^1_{cw}, Y^2_{cw}, \text{mask}).$$

Similarly, we can get the corresponding output of the other branch, which are denoted as $Y^1_{pw}$, $Y^2_{pw}$, and $Y_{ps}$, and the mixed output $Y_{pw}$. Instead of directly applying cross pseudo supervision between the predictions of the weak-augmented images and the strong-augmented images, e.g., $Y_{cw} \Rightarrow Y_{ps}$ and $Y_{cw} \Rightarrow Y_{cs}$, there are some additional operations applied to the $Y_{cw}$ and $Y_{pw}$ to generate the pseudo labels.

The pseudo label generation is based on the class-wise agreement and disagreement, and both of them are pixel-wise. For a pixel $p_i$, if the prediction of $Y_{cw}$ is the same as that of $Y_{pw}$, the two networks achieve agreement at $p_i$, otherwise they disagree with each other.

$$agreement = \begin{cases} \text{True} & y^i_{cw} = y^i_{pw} \\ \text{False} & \text{otherwise} \end{cases},$$

where $y^i_{cw}$ is the prediction of $Y_{cw}$ at $p_i$, and $y^i_{pw}$ is the corresponding prediction of $Y_{pw}$.

The pseudo label of a specific pixel is easy to determine, if the two predictions of the networks agree with each other. It is determined as the prediction at that pixel.

$$l^i_a = y^i_{cw}, \text{ if } y^i_{cw} = y^i_{pw},$$

where $l^i_a$ is the pseudo label for the agreement parts $L_a$ at the pixel $p_i$. However, it is much more complex to determine the pixel-wise pseudo labels for the disagreement parts $L_d$, as the predictions of the two networks are different. Instead of selecting the prediction with higher confidence, we determine the pseudo label based on the class-wise disagreement indicator, which is elaborated detailedly in the next subsection. Obtaining the pseudo labels for both the agreement and disagreement parts, the pseudo labels for union $L_{union}$ is compounded utilizing $L_a$ and $L_d$.

$$L_{union} = L_a \cup L_d.$$

Much simpler than that, the pseudo labels for intersection $L_{inter}$ is directly generated using the $L_a$.

Finally, the intersection and union pseudo supervision are applied to the conservative branch and progressive branch, respectively.

$$L_{inter} \Rightarrow Y_{cs},$$

$$L_{union} \Rightarrow Y_{ps},$$
The whole semi-supervised learning approach seeks common ground of the two paralleled branches while reserving differences, and achieves the collaboration of the conservative and the progressive.

Since the predictions of the two networks tend to agree with each other at the later training stage, which is indicated by the large overlap ratio (greater than 97%), we only preserve the network in conservative branch to generate segmentation predictions in inference mode.

C. Pseudo Label Determination for Disagreement

To deal with the disagreement between two predictions at the same pixel, a pseudo label determination approach for the disagreement part is introduced in this subsection, which is based on the class-wise disagreement indicator.

The typical and feasible approach is selecting the prediction with higher confidence as the pseudo label. However, this approach only focuses on the information of the exact pixel, which is limited. Addressing that, we turn to the class-wise disagreement indicator of the whole prediction to take a look from a macro perspective. Specifically, we construct the agreement matrix $M \in \mathbb{R}^{C \times C}$, as shown in Fig. 4, and calculate the class-wise disagreement indicator $I$ based on that. The semantic class with a higher $I$ is the difficult class, which is more likely to cause confusion. The pseudo label at the pixel $p_i$, where $y_{cw}^i \neq y_{pw}^i$, is determined as the prediction with higher disagreement indicator to achieve the progressive exploration for the disagreement part with strong curiosity. The pseudo label determination algorithm based on class-wise disagreement indicator is shown in Algorithm 1.

D. Confidence-Based Dynamic Loss

Benefit from the approaches proposed in the previous subsections, the pseudo labels of the unlabeled data are obtained. However, the noisy label is ineluctable, and there always be some errors in the generated pseudo labels since it is automatically annotated by the algorithm instead of human. Many researches have explored image denoising algorithms [41], [42]. Different from the image noise, the pseudo label noise is difficult to be either modeled or denoised. Addressing that, the confidence-based dynamic loss is proposed in this subsection to deal with the unreliable pseudo labels. Contrary to the confidence-thresholding, which requires a predefined threshold and could not take the full advantage of all the unlabeled data, we utilize the confidence for loss re-weighting.

The pixel-wise prediction confidence is defined as the maximum softmax probability. Let $b_c^i$ denotes the prediction confidence of the conservative branch at the $i$-th pixel, similarly, that of the progressive branch is denoted as $b_p^i$. Therefore, the dynamic confidence-based weight is defined as:

$$\omega_u^i = \begin{cases} \frac{1}{2}(b_c^i + b_p^i) & y_{cw}^i = y_{pw}^i \\ b_c^i & y_{cw}^i \neq y_{pw}^i, l_d^i \leftarrow y_{cw}^i \\ b_p^i & y_{cw}^i \neq y_{pw}^i, l_d^i \leftarrow y_{pw}^i. \end{cases}$$ (14)

Thus, the influence of the suspicious pseudo label with low confidence is reduced. The re-weighted loss on unlabeled data $\mathcal{L}_U$ is then defined as:

$$\mathcal{L}_U = \frac{1}{W \times H} \sum_{i=1}^{W \times H} \omega_u^i \cdot CE(l_{\text{inter}}^i, f_{b_c^i}(X_i)) + \frac{1}{W \times H} \sum_{i=1}^{W \times H} \omega_u^i \cdot CE(l_{\text{union}}^i, f_{b_p^i}(X_i)),$$ (15)

where $l_{\text{inter}}^i$ and $l_{\text{union}}^i$ are the pseudo labels of $L_{\text{inter}}$ and $L_{\text{union}}$ at the $i$-th pixel, the $f_{b_c^i}(X_i)$ is the output of the algorithm.
segmentation network at the $i$-th pixel for the unlabeled image $X_s$, and $CE(\cdot)$ is the cross-entropy loss function. The loss is re-weighted at each pixel according to the dynamic confidence-based weight.

The training leveraging the labeled data is in a traditional supervised manner.

$$L_S = CE(G, f_0(X_l)),$$

where $X_l \in D_l$ is the labeled image, and $G$ is the corresponding ground truth. The separated loss for supervised and unsupervised training is combined to the whole training objective, as

$$L = L_S + \gamma L_U,$$

where $\gamma$ is the trade-off weight to balance the two parts.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed approach CPCL on two public datasets, Cityscapes [2] and PASCAL VOC 2012 [43]. The experiments are implemented in the PyTorch on a server with a NVIDIA V100.

A. Implementation Detail and Dataset

The segmentation networks share the same structure, DeepLabv3+, and the weights of the segmentation heads are initialized randomly. The initial learning rate is set as $10^{-4}$ and $5 \times 10^{-3}$ when training with Cityscapes and PASCAL VOC, respectively, and the learning rate is multiplied by $(1 - \frac{\text{iter}}{\text{max}\_\text{iter}})^{0.9}$ in every iteration. The SGD with momentum is adopted. The momentum is fixed as 0.9, and the weight decay is $10^{-4}$. Due to the limitation of the GPU, the batch size for Cityscapes is set as 3, and that for PASCAL VOC is 8. The loss weight $\gamma$ is 5 and 1 respectively. We follow the partition protocols of GCT [39], which randomly extract 1/2, 1/4, 1/8, and 1/16 of the whole training set as the labeled subset, and regard the rest of the data as the unlabeled subset. If strong augmentation is to be applied, three rectangle regions of random ratio (with the range of $[0.25, 0.5]$) to the input is positioned randomly and augmented with CutMix strategy.

1) Cityscapes: is a dataset mainly designed for the urban driving scenes, which consists of 2975 images for training and 500 images for validation. Each image has a resolution of $2048 \times 1024$, and we random crop it to $800 \times 800$ with the original resolution kept. Each pixel is annotated with one of the semantic labels from 19 classes.
TABLE I
QUANTITATIVE RESULTS OF THE IMPROVEMENTS OVER THE SUPERVISED BASELINE ON CITYSCAPES. ALL SCORES ARE IN %, AND THE PERFORMANCE IMPROVEMENT IS SHOWN IN BRACKETS. (FLAT: ROAD AND SIDEWALK; HUMAN: PERSON AND RIDER; VEHICLE: CAR, TRUCK, BUS, TRAIN, MOTORCYCLE, AND BICYCLE; CONSTRUCTION: BUILDING, WALL, AND FENCE; OBJECT: POLE, TRAFFIC LIGHTS, AND TRAFFIC SIGN; NATURE: VEGETATION AND TERRAIN; SKY: SKY)

| Partition | Methods | mIoU | Flat | Human | Vehicle | Construction | Object | Nature | Sky |
|-----------|---------|------|------|-------|---------|--------------|--------|--------|-----|
| 1/2       | Supervised | 75.36 | 89.87 | 71.71 | 74.21 | 67.35 | 69.91 | 78.52 | 94.52 |
| 1/4       | CPCL     | 78.17 (2.81) | 91.44 (1.57) | 73.88 (2.16) | 79.29 (5.08) | 69.15 (1.80) | 72.44 (2.53) | 79.68 (1.16) | 94.72 (0.20) |
| 1/8       | Supervised | 73.22 | 88.88 | 71.53 | 70.40 | 64.07 | 69.80 | 75.98 | 94.34 |
| 1/16      | Supervised | 68.63 | 87.89 | 69.81 | 60.02 | 60.68 | 67.85 | 74.43 | 94.01 |
| 1/16      | CPCL     | 74.60 (5.97) | 89.48 (1.59) | 72.88 (3.07) | 72.89 (12.88) | 65.05 (4.36) | 70.49 (2.64) | 77.13 (2.70) | 94.40 (0.59) |
| 1/16      | Supervised | 61.67 | 85.27 | 65.66 | 46.88 | 53.49 | 63.03 | 72.71 | 93.51 |
| 1/16      | CPCL     | 69.92 (8.25) | 88.41 (3.15) | 70.15 (4.49) | 61.96 (15.08) | 62.69 (9.20) | 68.50 (5.47) | 75.97 (3.26) | 94.02 (0.51) |

TABLE II
QUANTITATIVE RESULTS OF THE IMPROVEMENTS OVER THE SUPERVISED BASELINE ON PASCAL VOC 2012. ALL SCORES ARE IN %, AND THE PERFORMANCE IMPROVEMENT IS SHOWN IN BRACKETS. (ANIMAL: BIRD, CAT, COW, DOG, HORSE, AND SHEEP; VEHICLE: AEROPLANE, BICYCLE, BOAT, BUS, CAR, MOTORBIKE, AND TRAIN; INDOOR: BOTTLE, CHAIR, DINING TABLE, POTTED PLANT, SOFA, AND MONITOR; PERSON; BACKGROUND: BACKGROUND)

| Partition | Methods | mIoU | Animal | Vehicle | Indoor | Person | Background |
|-----------|---------|------|--------|---------|--------|--------|------------|
| 1/2       | Supervised | 74.05 | 84.45 | 76.90 | 55.76 | 81.95 | 93.57 |
| 1/4       | CPCL     | 75.30 (1.25) | 86.63 (2.18) | 77.23 (0.33) | 57.11 (1.35) | 84.23 (2.28) | 94.12 (0.54) |
| 1/8       | Supervised | 71.66 | 78.98 | 74.31 | 55.91 | 82.05 | 93.23 |
| 1/16      | Supervised | 67.16 | 73.47 | 71.94 | 48.66 | 81.98 | 92.28 |
| 1/16      | CPCL     | 73.74 (6.58) | 84.24 (10.77) | 76.53 (4.59) | 54.91 (6.26) | 84.36 (2.38) | 93.60 (1.33) |

2) PASCAL VOC 2012: is a standard semantic segmentation dataset for common objects, which comprises 20 foreground classes along with a background class. The original dataset has 1446 images for training and 1449 validation images. We follow the previous works to generate the augmented version, which is combined with the extra annotation set from the Segmentation Boundaries Dataset (SBD) [44]. Thus, the full training set consists of 10582 images. We keep the original resolution and random crop the images to 512 × 512.

B. Effectiveness Study

Firstly, we evaluate our approach on Cityscapes to verify the effectiveness for the complex street scenes, and report the results on the validation set (500 images) for all partition protocols. The visualization examples on Cityscapes validation set are shown in Fig. 5. It can be seen from the first and the second rows that some pixels of road and sidewalk are missegmented by the supervised-only method. There are also confusion among train, truck and bus, as shown in the next four rows of the third column. Benefit from the proposed approach CPCL, the errors are corrected, and the predictions are more accurate compared with the supervised-only one.

The improvements over the supervised baseline on Cityscapes are significant, as shown in Fig. 6. Our approach consistently outperforms the DeepLabv3+ model under all partition protocols, and the gain of the CPCL increases as the ratio of the labeled data decreases. To further study the improvements, we report the quantitative results in Table I, which is more detailed. The performance increments are shown in the brackets. It can be seen that the gains of the approach are 2.81%, 3.76%, 5.97%, and 8.25% under 1/2, 1/4, 1/8, and 1/16 partition protocols separately. The utilization of the unlabeled data let the model achieve the similar performance as the one leverages more labeled data but only under supervised learning, e.g., the performance of CPCL under 1/16 partition is 69.92% and that of the supervised-only method under 1/8 partition is 68.63%. Compared with other class groups, the improvement for the vehicle is obvious, which achieves 15.08% increasement when only 186 labeled...
images are used. The gain on the sky class is limited since the supervised-only model has already achieved fine performance (about 94%) on that semantic class.

Different from Cityscapes, which focuses on traffic scenes, PASCAL VOC 2012 has richer scenes, and is the most commonly used benchmark for semi-supervised semantic segmentation. We evaluate our approach on it, and report the results on the validation set (1456 images) for all partition protocols.

In line with expectations, the CPCL consistently outperforms the supervised-only method under all partition protocols, and the gain is more obvious when fewer labeled images are used, as shown in Fig. 7. It can be seen from Table II that all of the metrics are increased due to the utilization of the unlabeled data. For different partitions, our approach achieves a performance increase of 2.81% (under 1/2 labels) to 8.25% (under 1/16 labels) over the baseline. The performance improvement of the animal class group is the most significant among the five groups, which goes up to 15.08% when only using 662 labeled images.

C. Ablation Study

The effectiveness of the proposed modules is further evaluated on Cityscapes under 1/8 partition protocol. We report the quantitative results in Table III. The intersection supervision and the union supervision are evaluated individually by training both the two paralleled networks under each of them. The cross-entropy loss function without re-weighting is used when the confidence-based dynamic loss is not adopted.

It can be seen that all of the three proposed modules are effective, resulting in the performance of the CPCL. First of all, the mIoU is improved from 68.63% to 71.57% when the intersection pseudo supervision is adopted. The improvement of 2.94% is achieved by leveraging a large number of unlabeled images. Since the intersection supervision only focuses on the agreement parts, the supervision is conservative and the improvement is limited. Different from that, the union pseudo supervision explores the disagreement parts with curiosity and leads to a more obvious improvement of 3.82% (from 68.63% to 72.45%). These two pseudo supervisions act in two different ways, conservative and progressive, and the collaboration of them improves the mIoU to 73.58%. Addressing the noisy label, we propose the confidence-based dynamic loss. The mIoU is increased by 2.37% and 1.70% as the dynamic loss is introduced. The improvement of the intersection supervision is more obvious, because the determination of the pseudo labels for the agreement parts is pixel-level, which is more likely influenced by the noisy label. Benefit from both the proposed pseudo supervision and the proposed dynamic loss, the mIoU of CPCL achieves 74.6%, overperforming the supervised baseline with 5.97%.

The proposed method is further compared with the mutual pseudo supervision, which directly utilizes the predictions to generate the pseudo labels and applies pseudo supervision based on that. The performances on the seven class groups are shown in Fig. 8. It’s worth noting that the performance of the direct mutual pseudo supervision is slightly worse than that of the supervised-only method on human and object groups. Those two groups consist of person, rider, traffic light, and traffic sign, which are small sized and complex. Thus, the influence of the erroneous pseudo supervision is obvious. Benefit from our pseudo labels generation method, the pseudo supervision is more reliable, resulting in the performance improvements on all the groups.

Finally, we study different pseudo labeling strategies for the disagreement part and analyze the quantitative results. Four related strategies are compared, including the pseudo
labeling on the basis of the pixel-wise confidence (higher and lower) and class-wise confusion (higher and lower), and the results are reported in Table IV. The pixel-wise confidence indicates the uncertainty of the exact pixel. Pseudo labeling with higher confidence is a common conservative strategy. However, the improvement is less obvious when collaborating with the intersection pseudo supervision, for both of them act conservatively. Exploration on the basis of lower confidence is progressive, but it is more likely influenced by pixel-wise noise. The proposed dynamic loss can reduce the influence to some extent, and the mIoU of it achieves 74.30%. Different from the pixel-wise confidence, the class-wise confusion can measure the semantic ambiguity from a macro perspective. Pseudo labeling with lower confusion is similar to that based on higher confidence, but at class-level, which has the worst performance (74.17% mIoU) among the four compared strategies. Since exploring the difficult semantic class with strong curiosity may help the models out the local optimization, the proposed strategy used in CPCL achieves a better performance.

**D. Comparison With Others**

We compare our approach with other semi-supervised segmentation methods in this subsection. All compared methods are based on DeepLabv3+ with ResNet-50 or ResNet-101, and the reported results of others are taken from [35].

Firstly, the quantitative results on Cityscapes are compared. As shown in Table V, CPCL outperforms all reported methods with ResNet-50, and achieves similar or even better performance when compared with others with ResNet-101. In addition, it can be seen that the performance advantage becomes more evident when leveraging fewer labeled images for training.

Then, the comparison between the CPCL and others on PASCAL VOC 2012 is reported in Table VI. Due to the limitation of the GPU, the batch size is set as 7 when evaluating the CPCL with ResNet-101. Similar to that on Cityscapes, the performance advantage of our approach is more evident when using fewer labeled images. As the ratio of labeled data increases, the advantage decreases. However, CPCL with ResNet-101 still achieves state-of-the-art performance and outperforms the compared methods. Moreover, the CPCL with ResNet-50 even performs better than those with ResNet-101 under the partitions 1/8 labels.

![Graph](image_url)  
**Fig. 9.** Comparison for full-supervision on PASCAL VOC 2012. We use the full PASCAL VOC 2012 training set (1.4k) as the labeled data, and the images from SBD (9k) as the additional unlabeled data.

In addition, the proposed framework CPCL does not restrict the backbones. The two parallel branches can use two different backbones. For example, one with ResNet-50 and another with ResNet-101. The experiments are further conducted on PASCAL VOC under 1/8 partition protocol. If the backbone of the conservative branch is changed to ResNet-101, the mIoU is improved to 73.94%. The improvement is more obvious when ResNet-101 is used in the progressive branch, which achieves 74.85% mIoU. Note that the reported result is the better one of the two parallel branches, and it is usually output from the progressive branch.

**E. Improving Full-Supervision With Additional Data**

The improvement over the full-supervision of our approach is verified in this subsection. The full PASCAL VOC 2012 training set (1.4k) is used as the labeled data, and the images from SBD (9k) are utilized as the additional unlabeled data. We compare the proposed approach with recent state-of-the-arts methods, as shown in Fig. 9.

All the results of others are taken from [37]. The backbone of GANSeg [26] is VGG16, while that of the CCT [10] is...
TABLE VII
Quantitative Results of Different Methods on PASCAL VOC 2012 in Low-Data Regime. The Metric Is mIoU(%), and the Results of the Compared Methods Are From [37]

| Methods            | Network | Exact Numbers of Labeled Data |
|--------------------|---------|-------------------------------|
|                    |         | 92   | 183  | 366  | 732  |
| CCT [10]           | ResNet-50 | 33.10 | 47.60 | 58.80 | 62.10 |
| VAT [7]            | ResNet-101 | 36.92 | 49.35 | 56.88 | 63.34 |
| AdvSemSeg [5]      | ResNet-101 | 39.69 | 47.58 | 59.97 | 65.27 |
| GCT [39]           | ResNet-101 | 46.04 | 54.98 | 64.71 | 70.67 |
| MT [6]             | ResNet-101 | 48.70 | 55.81 | 63.01 | 69.16 |
| CutMix-Seg [9]     | ResNet-101 | 55.58 | 63.20 | 68.36 | 69.84 |
| PseudoSeg [37]     | ResNet-101 | 57.60 | 65.50 | 69.14 | 72.41 |
| CPCL (Ours)        | ResNet-50  | 61.88 | 67.02 | 72.14 | 74.25 |

ResNet-50, which is the same as ours. Both the AdvSemSeg [5] and the PseudoSeg [37] use ResNet-101 as the backbone. It can be seen that CPCL achieves the best performance among all the compared methods. The mIoU of our approach is 76.47%, which is better than the GANSeg with a large margin (about 12%). The second higher method, PseudoSeg, is still 3.24% lower than ours.

F. Improving Few-Supervision in Low-Data Regime

We further evaluate our approach in low-data regime on PASCAL VOC 2012. The partition protocols in PseudoSeg [37] is adopted in experiment, in which the labeled data set is constructed by sampling part of the images in the standard training set. The unlabeled set is generated by utilizing the additional unlabeled data in the SBD and the remaining images in the standard training set. The comparison between CPCL and other methods is reported in Table VII.

The performance advantage of our approach in low-data regime is obvious. Our approach performs best among all the compared methods. It is superior to CCT [10], which also uses ResNet-50 as the backbone, and the improvement under all partition protocols are more than 10%, from 12.15% (using 732 labels) to 28.78% (using 92 labels). Compared with other methods utilizing ResNet-101 as the backbone, CPCL still shows a significant performance improvement. Among the methods with ResNet-101, PseudoSeg [37] has the best overall performance, while CPCL achieves the increments of 1.84%, 3.00%, 1.52%, and 4.28% under different partition protocols. In addition, it can be seen that the mIoU (72.14%) of our approach using 366 labeled images is comparable with that (73.23%) of PseudoSeg using the full training set. The demand for labeled data can be effectively reduced.

V. CONCLUSION AND DISCUSSION

In this paper, we propose a novel semi-supervised learning approach for semantic segmentation, termed Conservative-Progressive Collaborative Learning (CPCL), to not only take the advantage of the high-quality labels but also make the full use of the large quantity of the unlabeled data. We realize the CPCL by the intersection and union pseudo supervision. The intersection pseudo supervision based on the agreement of the two paralleled networks is conservative and reliable, while the union pseudo supervision achieves the progressive exploration for the disagreement parts. They cooperate with each other and achieve the collaboration of the conservative evolution and the progressive exploration. To reduce the pseudo supervision noise, the confidence-based dynamic loss is proposed, which re-weights the loss on the basis of the prediction confidence. CPCL is simple, efficient and flexible. We have evaluated the proposed approach on two widely used datasets, Cityscapes and PASCAL VOC 2012. The extensive experimental results demonstrate the effectiveness of the CPCL, and show its state-of-the-art performance for semi-supervised semantic segmentation, especially in low-data regime.

Although the performance is effectively improved, error predictions are unavoidable. It can be seen from the fourth row in Fig. 5 that there are still segmentation errors between terrain and sidewalk. Union pseudo supervision is adopted for exploration, but the effect is limited when both the two networks are confident with wrong predictions. In our opinion, the future works can be developed from the following perspectives. Enhancing the quality of the pseudo supervision is the key to the good performance for the semi-supervised semantic segmentation. However, it is also a complicated and difficult task. We have tried several approaches, and most of them are mediocre. The main problem is how to estimate the reliability of the pseudo supervision without ground truth labels. In this paper, we leverage the agreement of the two paralleled models and the prediction confidence of the single model, but it is insufficient if model learned on labeled data is insufficient. Therefore, the evaluation of the pseudo supervision deserves further studied. Besides, the model coupling is another problem, which causes the peer-supervised networks to converge towards a wrong direction. Heterogeneous architecture is an effective way to reduce that. In this paper, we applied intersection pseudo supervision and union pseudo supervision to the two branches respectively. The status of the model coupling is usually empirically observed from the overlap ratio between the two models during training, which is usually small at the early training stage and increases during the later training. However, the overlap ratio may not be an effective indicator for model coupling, since we cannot distinguish the model coupling problem from the accurate models reaching a high degree of consensus with each other. Thus, the indicator for the model coupling problem is also worth to be further study.

REFERENCES

[1] F. Zhu, Y. Lv, Y. Chen, X. Wang, and F. Wang, “Parallel transportation systems: Toward IoT-enabled smart urban traffic control and management,” IEEE Trans. Intell. Transp. Syst., vol. 21, no. 10, pp. 4063–4071, Oct. 2020.
[2] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 3213–3223.
[3] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.
[4] Y. Grandvalet and Y. Bengio, “Semi-supervised learning by entropy minimization,” in Proc. Adv. Neural Inf. Process. Syst., 2005, pp. 529–536.
[5] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, “Adversarial learning for semi-supervised semantic segmentation,” 2018, arXiv:1802.07934.

[6] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 1196–1205.

[7] T. Miyato, S.-L. Maeda, M. Koyama, and S. Ishii, “Virtual adversarial training: A regularization method for supervised and semi-supervised learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 8, pp. 1799–1803, Aug. 2019.

[8] X. Wang, D. Kihara, J. Luo, and G.-J. Qi, “EnAET: A self-trained framework for semi-supervised and supervised learning with ensemble transformations,” IEEE Trans. Image Process., vol. 30, pp. 1639–1647, Oct. 2020.

[9] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, “Semi-supervised semantic segmentation needs strong, varied perturbations,” 2019, arXiv:1906.01916.

[10] Y. Ouali, C. Hudelet, and M. Tami, “Semi-supervised semantic segmentation with cross-consistency learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12671–12681.

[11] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. Lau, “Dual student: Breaking the limits of the teacher in semi-supervised learning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6727–6735.

[12] E. Arazo, D. Ortego, P. Albert, N. E. O’Connor, and K. McGuinness, “Pseudo-labeling and confirmation bias in deep semi-supervised learning,” in Proc. Int. Joint Conf. Neural Inf. Netw. (IJCNN), Jul. 2020, p. 1.

[13] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, “In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning,” 2021, arXiv:2101.06329.

[14] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, “Deep co-training for semi-supervised image recognition,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 142–159.

[15] Z. Feng et al., “DMT: Dynamic mutual training for semi-supervised learning,” 2020, arXiv:2004.08514.

[16] P. Zhang, B. Zhang, T. Zhang, D. Chen, and F. Wen, “Robust mutual learning for semi-supervised semantic segmentation,” 2021, arXiv:2106.00609.

[17] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in Proc. Workshop Challenges Represent. Learn. (ICML), 2013, p. 896.

[18] Y. Zou, Z. Yu, B. V. K. V. Kumar, and J. Wang, “Unsupervised domain adaptation for semantic segmentation via class-balanced self-training,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 297–313.

[19] L.-C. Chen et al., “Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 695–714.

[20] B. Zoph et al., “Rethinking pre-training and self-training,” in Proc. Adv. Neural Inf. Process. Syst., 2020, p. 1.

[21] M. S. Ibrahim, A. Vahdat, M. Ranjabar, and W. G. Macready, “Semi-supervised semantic image segmentation with self-correcting networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12712–12722.

[22] R. Mendel, L. A. D. Souza, D. Rauber, J. P. Papa, and C. Palm, “Semi-supervised segmentation based on error-correcting supervision,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 141–157.

[23] C. Li, T. Xu, J. Zha, and B. Zhang, “Triple generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4089–4099.

[24] Y. Luo, J. Zha, M. Li, Y. Ren, and B. Zhang, “Smooth neighbors on teacher graphs for semi-supervised learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8896–8905.

[25] J. Tobias Springenberg, “Unsupervised and semi-supervised learning with categorical generative adversarial networks,” 2015, arXiv:1511.06390.

[26] N. Souly, C. Spampinato, and M. Shah, “Semi-supervised semantic segmentation using generative adversarial network,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5689–5697.

[27] S. Mittal, M. Tataruchenko, and T. Brox, “Semi-supervised semantic segmentation with high- and low-level consistency,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 4, pp. 1360–1379, Apr. 2021.

[28] B. Athiwaratkun, M. Finzi, P. Izmailov, and A. G. Wilson, “There are many consistent explanations of unlabeled data: Why you should average,” in Proc. Int. Conf. Learn. Represent., 2018, p. 1.

[29] S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning,” in Proc. Int. Conf. Learn. Represent., 2017, p. 1.
Zunlei Feng received the B.Eng. degree from Soochow University and the Ph.D. degree in computer science and technology from the College of Computer Science, Zhejiang University. He is an Assistant Research Fellow with the College of Software Technology, Zhejiang University. He has authored and coauthored many scientific articles at top venues, including *International Journal of Computer Vision* (IJCV), NeurIPS, AAAI, *IEEE Transactions on Visualization and Computer Graphics* (TVCG), *ACM TOMM*, and ECCV. He has served with international conferences, including NeurIPS, ICCV, ICML, IJCAI, AAAI, and PKDD, and international journals, including *IEEE Transactions on Circuits and Systems for Video Technology*, *Information Sciences*, *Neurocomputing*, *Journal of Visual Communication and Image Representation*, and *Neural Processing Letters*. His research interests mainly include computer vision, image information processing, representation learning, and medical image analysis.

Yisheng Lv (Senior Member, IEEE) received the B.E. and M.E. degrees from the Harbin Institute of Technology, Harbin, China, in 2005 and 2007, respectively, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2010. He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences. His current research interests include traffic data analysis, dynamic traffic modeling, and parallel traffic management and control systems.

Mingli Song (Senior Member, IEEE) received the Ph.D. degree in computer science from Zhejiang University, China, in 2006. He is currently a Professor with the Microsoft Visual Perception Laboratory, Zhejiang University. His research interests include face modeling and facial expression analysis. He received the Microsoft Research Fellowship in 2004.

Fei-Yue Wang (Member, IEEE) received the Ph.D. degree in computer and systems engineering from the Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. In 1990, he joined the University of Arizona and became a Professor and the Director of the Robotics and Automation Laboratory (RAL) and Program in Advanced Research for Complex Systems (PARCS). In 1999, he founded the Intelligent Control and Systems Engineering Center at the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Oversea Chinese Talents Program from the State Planning Council and “100 Talent Program” from CAS, and in 2002, was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory of Management and Control of Complex Systems. His current research interests include methods and applications for parallel systems, social computing, and knowledge automation. Since 2008, he has been the Vice President and the Secretary General of Chinese Association of Automation. He is an Elected Fellow of IEEE, INCOSE, IFAC, ASME, and AAAS. He was the President of IEEE ITS Society from 2005 to 2007, the Chinese Association for Science and Technology (CAST, USA) in 2005, the American Zhu Kezhen Education Foundation from 2007 to 2008, and the Vice President of the ACM China Council from 2010 to 2011. In 2007, he received the 2nd Class National Prize in Natural Sciences of China and awarded the Outstanding Scientist by ACM for his work in intelligent control and social computing. He received IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, and the IEEE SMC Norbert Wiener Award in 2014. Since 1997, he has been the General or Program Chair of more than 20 IEEE, INFORMS, ACM, and ASME conferences. He was the Founding Editor-in-Chief of the *International Journal of Intelligent Control and Systems* from 1995 to 2000 and *IEEE Intelligent Transportation Systems Magazine* from 2006 to 2007. He was the Editor-in-Chief of *IEEE Intelligent Systems* from 2009 to 2012 and *IEEE Transactions on Intelligent Transportation Systems* from 2009 to 2016. Currently, he is the Editor-in-Chief of China’s *Journal of Command and Control*.