Adversarial Dual-Student With Differentiable Spatial Warping for Semi-Supervised Semantic Segmentation

Cong Cao, Tianwei Lin, Dongliang He, Fu Li, Huanjing Yue, Member, IEEE, Jingyu Yang, Senior Member, IEEE, and Errui Ding

Abstract—A common challenge posed to robust semantic segmentation is the expensive data annotation cost. Existing semi-supervised solutions show great potential for solving this problem. Their key idea is constructing consistency regularization with unsupervised data augmentation from unlabeled data for model training. The perturbations for unlabeled data enable the consistency training loss, which benefits semi-supervised semantic segmentation. However, these perturbations destroy image context and introduce unnatural boundaries, which is harmful for semantic segmentation. Besides, the widely adopted semi-supervised learning framework, i.e., mean-teacher, suffers performance limitation since the student model finally converges to the teacher model. In this paper, first of all, we propose a context friendly differentiable geometric warping to conduct unsupervised data augmentation; secondly, a novel adversarial dual-student framework is proposed to improve the Mean-Teacher from the following two aspects: (1) dual student models are learned independently except for a stabilization constraint to encourage exploiting model diversities; (2) adversarial training scheme is applied to both students and the discriminators are resorted to distinguish reliable pseudo-label of unlabeled data for self-training. Effectiveness is validated via extensive experiments on PASCAL VOC2012 and Cityscapes. Our solution significantly improves the performance and state-of-the-art results are achieved on both datasets. Remarkably, compared with fully supervision, our solution achieves comparable mIoU of 73.4% using only 12.5% annotated data on PASCAL VOC2012. Our codes and models are available at https://github.com/cao-cong/ADS-SemiSeg.

Index Terms—Semi-supervised semantic segmentation, dual-student, differentiable spatial warping.

I. INTRODUCTION

Understanding images at pixel level is prevalent due to that it enables many practical applications such as scene parsing, environment sensing for auto driving or visual navigation. Therefore, semantic segmentation is becoming a more and more important computer vision task, which aims at semantically labeling natural images on per-pixel basis. Recently, owing to the great progress in deep learning, numerous CNN frameworks have been developed for this task [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. However, the data driven training paradigm of CNNs requires a huge amount of manually labeled data. Especially for semantic segmentation, the annotations must be accurate at pixel level, leading to expensive cost for annotating an image. Therefore, data annotation poses great challenge for us to further improve the segmentation performance efficiently.

Reviewing the literature, semi-supervised learning, which attempts to leverage unlabeled data to help model training, is widely utilized to relieve annotation burden by the computer vision community. Many frameworks have been specially designed for semi-supervised semantic segmentation [11], [12], [13]. A key trend of previous work is to apply unsupervised data augmentations, such as CutMix [11] and CowMix [12], to images and then construct consistency constraint. Although these strategies are able to improve the semi-supervised segmentation performance, they destroy the image context and introduce weird boundaries. The recently proposed ClassMix [13] achieves better performance since it replaces the weird boundaries by the class boundaries, as shown in Fig. 1. However, it still causes severe scene context destruction, harming semantic segmentation models.
The other widely utilized practice is to leverage the well-known state-of-the-art semi-supervised learning framework called mean-teacher [14] and adapt it to semantic segmentation scenario for training under the consistency regularization. However, mean-teacher ends up with a student model which is an exponential moving average version of the teacher model. Such moving average will make the student finally converge to the teacher model and therefore model diversity is not fully exploited. Under the mean-teacher framework, the performance of student model will be limited by the teacher [15].

**To relieve the scene context destruction**, we propose a Differentiable Geometric Warping (DGW) mechanism to perform image perturbation. As can be seen from Fig. 1, our DGW distorts image content spatially and can better preserve semantic context compared to recent state-of-the-art CowMix [12] and ClassMix [13]. For both rigid and non-rigid objects, CowMix and ClassMix will result in disharmonious surroundings around the object contours, e.g. dogs over the sky and airplane near dogs. Context is proven to be of great importance for segmentation task [3]. Nevertheless, such discordant context caused by CowMix or ClassMix is not the case for real natural images, so it can be harmful for segmentation models. With our DGW, though the shape of rigid and non-rigid objects will be slightly distorted, it is still recognizable. Besides, the context near object contours is still natural. Therefore, for a segmentation model, correctly assigning class labels and precisely determining object boundaries for spatially geometric warped images are still possible. We also empirically verify that DGW achieves better performance than several commonly used image perturbations. In addition, it is worthy of noting that forcing models to predict consistency segmentation maps with respect to known spatial transformations is essential for stabilized model prediction. Under the DGW setting, the extra supervision signal then becomes that, the prediction of distorted image should be consistent with the corresponding distorted prediction of the original image. Specifically, in this paper, DGW is implemented based on Thin Plate Spline (TPS) [16] algorithm, which is a differentiable operation and fits a mapping from source to destination image coordinates.

**To alleviate the drawback of mean-teacher**, in this paper, we propose to reform the recent state-of-the-art semi-supervised image classification framework named dual-student [15] to an adversarial setting for semi-supervised semantic segmentation. Our adversarial dual-student framework can: (1) better utilize unlabeled image data and (2) enable to exploit model diversity. In more detail, two student networks are designed to learn from both supervised signal provided by labeled data and consistency regularization constructed with unlabeled data. The two students are relatively independent from each other and only reliable knowledge of both students can be exchanged. In this way, diversity (or variation) can be introduced and neither of the student model would be limited by the other one. Besides, adversarial training scheme is used for each student. For both student models, discriminators are trained to distinguish reliable predictions for unlabeled images and such reliable predictions are then leveraged as self-training signal to further improve our model.

In a nutshell, we focus on pushing the limitation of semi-supervised semantic segmentation and our contributions can be summarized as follows:

- We propose a differentiable geometric warping mechanism, which can be easily implemented with TPS and is context friendly, as unsupervised image augmentation for semi-supervised semantic segmentation.
- We are the first to reform dual-student framework to semantic segmentation task, and further propose the Adversarial Dual-Student (ADS) to benefit semi-supervised training from self-training mechanism.
- Extensive experiments show that our solution significantly improve the performance and new state-of-the-art performance is achieved on both PASCAL VOC2012 and Cityscapes datasets. Remarkably, compared with fully supervision, our solution achieves comparable mIoU of 73.4% using only 12.5% annotated data on PASCAL VOC2012.

## II. RELATED WORKS

### A. Semi-Supervised Learning

Semi-supervised learning is a classic topic in both computer vision and machine learning industry, which is recently drawing much attention in image classification task. In a major line, consistency regularization is used to take unlabeled data into account for defining the decision boundaries, thus can exploit unlabeled data to improve results [14], [17], [18], [19], [20]. Typically, mean-teacher [14] proposes a teacher-student framework, which applies consistency-regularization between a student and a teacher based on Exponential Moving Average (EMA) of students in each update step. Recently, Dual-Student [15] train two independent student networks and apply stabilization constraint instead of consistency constraint, avoiding performance bottleneck caused by coupled EMA teacher. Another way is to predict for unlabeled data with a network and use the predictions as pseudo labels [14], [21], [22].

### B. Semi-Supervised Semantic Segmentation

Compared with image classification, semi-supervised semantic segmentation is far more difficult and complex and also has two main lines. One line exploits consistency regularization, [11], [12], [13] focuses on different perturbation strategies, [23], [24], [25] focuses on the architecture of consistency training. The other line aims to encourage network to make confident predictions on unlabeled data. The work in [26] and [27] uses an additional discriminator to select real enough predictions as pseudo labels. The work in [28], [29], and [30] uses dynamic self-training strategy to improve quality of pseudo labels. The work in [31] uses an correction network to correct segmentation networks prediction. The works in [32] and [33] modify the segmentation network with self-attention module and apply spectral normalization to the GAN discriminator to improve the stability of the semi-supervised training. The work in [34] proposes a two-branch network to encode strong and pseudo label spaces respectively, which could well utilize pseudo-labels. Besides learning from pseudo
In this work, we combine the two lines via proposing an Adversarial Dual-Student framework. We design a segmentation stabilization loss helps them to exchange reliable knowledge. The overall proposed framework is illustrated in Fig. 2. The ADS framework contains two exactly the same student branches without weight sharing. Within each branch, labeled data contributes to supervised training loss, while the unlabeled data is used by consistency constraints. Between two branches, stabilization loss helps them to exchange reliable knowledge.

C. Perturbation Strategies

In consistency regularization, predictions of unlabeled data are required to be invariant to perturbations. Recently, a lot of data augmentation methods [11], [12], [13], [48], [49], [50], [51], [52], [53] were proposed and some of them were applied to produce perturbations. Among them, CutMix [11], CowMix [12], and ClassMix [13] are mask-based mixing method which cut out regions from one image and paste onto another. Recently, ClassMix [13] combines images using binary masks generated upon segmentation results instead of random rectangles and achieves state-of-the-art performance. However, this kind of mixing usually leads to context distortion, generating unreasonable training pairs. Therefore, instead of mask-based mixing, in this work, we suggest that segmentation consistency can be treated as a equivalent constraint problem [54], [55], which requires that prediction is invariant to spatial warping and thus leads to robust prediction. To achieve this, we propose a differentiable spatial warping layer based on TPS.

III. APPROACH

A. Overview

1) Problem Formulation: Given an image \( x \in \mathbb{R}^{H \times W \times 3} \), it can be annotated with a semantic segmentation map \( y \in \mathbb{R}^{H \times W} \), where \( y_{i,j} \in \{0, 1, \ldots, c\} \) is the semantic label of pixel \((i, j)\) and \(c\) is the number of target categories. In semi-supervised learning, for a full image set \( X = [X_l, X_u] \), annotations are only provided in labeled subset \( X_l \), unlabeled subset is denoted by \( X_u \). Our target is to train a semantic segmentation model with images from both labeled and unlabeled subset.

2) Overall Framework: The overall proposed framework is illustrated in Fig. 2. The ADS framework contains two exactly the same student branches without weight sharing. Within each branch, labeled data contributes to supervised training loss, while the unlabeled data is used by consistency constraints. Both \( x_u \) and its warped version \( T(x_u) \) are predicted via a student \( S \) and we require that \( T(S(x_u)) \) and \( S(T(x_u)) \) should be well aligned. Adversarial learning is also equipped for both students. Adversarial training can result in a discriminator to distinguish high-quality pseudo labels for unlabeled data and such pseudo labels can be leveraged to improve the model training. Between two branches, stabilization loss helps them exchange reliable knowledge.

B. Differentiable Geometric Warping

In the wild, objects may have various locations, poses and orientations, bringing great challenge to semantic segmentation. On the other hand, rich potential context also exists in these natural image deformations. For example, in unsupervised keypoints detection methods [54], [55], forcing model to predict consistency keypoints with respect to known geometric transformations is the key factor for discovering stable keypoints. Thus, how can we exploit these potential context in a controlled way to improve semantic segmentation? We believe the semantic segmentation of objects should also be invariant to controlled geometric transformation, which can also lead to more stable segmentation results. Besides, the geometric transformation should meet the requirements of (1) objects in perturbed images should be recognizable for correct label assigning and (2) the contextual information near object contours should be natural for precise boundary determination. To achieve this, we adopt TPS [16] deformation to construct our consistency regularization mechanism, which...
has following advantages: (1) can achieve more complex deformation than basic affine transform and be similar to where warping is controlled by nine random source (red)-destination (green) point-pairs and four fixed ones.

Based on TPS, we propose the DGW layer as a new consistency regularization mechanism. Fig. 3 illustrates how DGW works. For an image \( x \in \mathbb{R}^{h \times w \times 3} \), we equally split it into \( n \times n \) blocks. Then, for the \( i \)-th block, we can obtain a source point as \( p_i^s = (u_i, \delta^s_u, v_i, \delta^s_v) \), where \((u_i, v_i)\) is the center point of block and \( \delta^s \sim N(0, \sigma_d) \) is a Gaussian function. Further, we calculate corresponding destination point \( p_i^d = (u_i + \delta^d_u, v_i + \delta^d_v + \delta^d_t) \), where \( \delta^d \sim N(0, \sigma_d) \).

Finally, \( P_S \) and \( P_D \) are used to control the TPS transformation to generate warped image \( T(x) \), where the deformation degree is controlled by \( \sigma_d \).

C. Adversarial Dual-Student Framework

In this section, we will introduce the details of our proposed ADS framework. First, we introduce how we reform the Dual-Student framework for semantic segmentation task based on DGW. Then, the details of adversarial training are introduced. Fig. 2 presents the schematic diagram of our ADS framework.

1) Dual-Student: Most existing semi-supervised semantic segmentation methods [11, 47] adopt mean-teacher framework. However, as discussed in [15], because of exponential moving average, teacher network in mean-teacher tends to be very close to the student when the training process converges, causing the lacking of diversity and performance bottleneck. To overcome this coupled teacher problem, Dual-Student [15] is proposed to use two independently initialized student network without teacher. The core idea of Dual-Student is applying a stabilization constraint to unlabeled data to exchange reliable knowledge. In this work, we propose a novel consistency constraint and a novel stabilization constraint for semantic segmentation based on Dual-Student. We denote the two student segmentation networks as \( S_a \) and \( S_b \) respectively. In the following, we introduce our ADS framework via introducing each loss function. For simplicity, we only show loss functions of \( S_a \). The losses for the other student model \( S_b \) can be analogically calculated.

2) Cross-Entropy Loss: For labeled samples \( x_l \), we directly apply supervised pixel-level cross entropy loss function \( L_{ce} \) for each network. With the network output \( S_a(x) \in \mathbb{R}^{h \times w \times C} \), we can calculate cross-entropy loss as:

\[
L_{ce}^a = -\sum_{i,j,c} y_{(i,j,c)} \log S_a(x_i)(i,j,c)
\]

3) Consistency Constraint: To exploit context of unlabeled samples \( x_u \), we apply consistency constraint based on DGW transform for each branch separately. For an unlabeled image \( x_u \), segmentation network \( S_a \) is applied twice, where DGW transform with the same control points is applied before and after \( S \) separately. With two generated results \( S_a(T(x_u)) \) and \( T(S_a(x_u)) \), we can calculate a pixel-wise consistency error \( e^a \in \mathbb{R}^{h \times w} \), which is used for calculating both consistency and stabilization loss:

\[
e^a(x_u)(i,j) = (S_a(T(x_u))(i,j) - T(S_a(x_u))(i,j))^2
\]

This consistency constraint encourages each student network to be invariant to geometric transform, leading network to learn how to handle natural objects with complex shapes, orientations, and deformations.

4) Stabilization Constraint: The core idea of Dual-Student [15] is that only reliable knowledge could be exchanged between two student networks. To achieve this, Dual-Student first defines stable samples, and then elaborates the derived stabilization constraint to exchange knowledge. In this work, we further extend the concept of stable samples and stabilization constraint from image-level to pixel-level to achieve semi-supervised semantic segmentation.

What is a stable pixel in semantic segmentation? We suggest that a stable pixel should be invariant to spatial perturbation. Thus, for each student network \( S \), a stable pixel should meet following requirements: (1) semantically the same: the predicted semantic label of \( S_a(T(x))(i,j) \) and \( T(S_a(x))(i,j) \) should be the same; (2) high confidence: either of \( S_a(T(x))(i,j) \) and \( T(S_a(x))(i,j) \) should be larger than a pre-defined threshold \( \zeta \). If a pixel \((i, j)\) satisfies these two conditions at student \( A \), we denote stable score \( r^a_{(i,j)} = 1 \), otherwise 0. Then, we can derive the stabilization constraint in different situations: (1) if a pixel is not stable, no constraint will be applied; (2) if a pixel is stable for one student but unstable for the other, the stable one is used to constrain the unstable one. (3) if a pixel is stable for both student, then we use the more stable one to constrain the other one. Thus, for student \( A \), we can calculate pixel-wise stabilization loss as:

\[
l_{sta}^a(x) = \begin{cases} [e^a(x) < e^b(x)]_1 L_{mse}(x) & r^a = r^b = 1 \\ r^a L_{mse}(x) & otherwise \end{cases}
\]
where \(|l|\) is an indicator and it equals 1 only when student A is more stable, i.e., \(e^a(x) < e^b(x)\). Otherwise, it is 0. 

\(L_{mse}\) measures the mean square error between outputs of two student networks:

\[
L_{mse}(y_a(i,j)) = (S_a(T(x_a))(i,j) - S_b(T(x_a))(i,j))^2
\]

For simplicity, we omit the pixel position \((i, j)\) in Eq. 4. Finally, the total stabilization loss can be calculated as

\[
L_{sta}(x_a) = \frac{1}{h \times w} \sum_{i,j} f_{sta}^a(x_a)(i,j)
\]

The proposed stabilization loss can effectively exchange reliable knowledge between two segmentation networks, meanwhile keeping the diversity.

5) Adversarial Learning: To further exploit knowledge in unlabeled data, we combine conditional adversarial learning with our dual-student framework. Specifically, we adopt the discriminator of s4GAN [47] as the adversarial module in each branch of Dual-Student. Formally, an extra discriminator \(D_a\) needs to be trained for \(S_a\) with original GAN loss as:

\[
L_{D}^a = \mathbb{E}[\log D_a(y_l \oplus x_l)] + \mathbb{E}[\log(1 - D_a(S_a(x_a) \oplus x_a))]
\]

where \(\oplus\) denotes concatenation along channels. Based on discriminator, we can calculate feature matching loss to minimize the mean discrepancy of the feature statistics between \(S_a(x_a)\) and \(y_l\):

\[
L_{fm}^a = ||\mathbb{E}[D_a^{(k)}(y_l \oplus x_l)] - \mathbb{E}[D_a^{(k)}(S_a(x_a) \oplus x_a)]]||
\]

where \(D_a^{(k)}\) denotes the \(k\)-th intermediate feature map of \(D_a\). Furthermore, following [47], we adopt self-training loss \(L_{st}^a\) to pick the well predicted pixels (which can fool \(D_a\) and are predicted as “Real” with score larger than some threshold) for supervising \(S_a\). The loss term \(L_{st}^a\) is defined as:

\[
L_{st}^a = \begin{cases} 
-\sum_{h,w,c} \hat{y} \log \hat{s}_a(x_u), & \text{if } D_a(S_a(x_u) \oplus x_u) \geq \gamma, \\
0, & \text{otherwise}.
\end{cases}
\]

where \(\hat{y}\) denotes pseudo pixel-wise labels generated from segmentation result \(S_a(x_u)\) and \(\gamma\) is a threshold that determines how confident about the prediction that \(D_a\) needs to be. In our implementation, \(\gamma\) is set to be 0.6 and 0.9 for PASCAL VOC2012 and Cityscapes, respectively. \(\hat{y}\) is generated by sharpen the prediction value from continuous real values to be discrete value in \([0, 1]\) by converting logit to one-hot pseudo-label. We denote the overall adversarial loss as:

\[
L_{adv}^a = \lambda_{fm} \cdot L_{fm}^a + \lambda_{st} \cdot L_{st}^a + \lambda_{st} \cdot L_{sta}^a
\]

where \(\lambda_{fm}, \lambda_{st}\) are weighting factors.

6) Training Objective: With aforementioned loss functions, for student model \(S_a\), the loss is combined as:

\[
L_a = L_{ce}^a + \lambda_1 \cdot L_{cons}^a + \lambda_2 \cdot L_{sta}^a + \lambda_3 \cdot L_{adv}^a
\]

The overall training objective is to minimize the following loss function \(L = L^a + L^b\).
TABLE I
EVALUATION OF THE EFFECTIVENESS OF EACH COMPONENT ON PASCAL VOC2012 AND CITYSCAPES. WE REPORT MIoUs WHEN ONLY 1/8 AND 1/4 TRAINING SAMPLES ARE LABELED. MT, DS AND ADV DENOTE MEAN-TEACHER, DUAL-STUDENT AND ADVERSARIAL TRAINING ($L_{fm} + L_D + L_{st}$), RESPECTIVELY.

| Dataset   | VOC2012 | Cityscapes |
|-----------|---------|------------|
| Method    | 1/8     | 1/4        | 1/8     | 1/4     |
| Supervised DeepLab-v2 | 65.2 | 68.9 | 55.0 | 60.3 |
| MT + DGW  | 70.6    | 73.2       | 60.1    | 63.2   |
| MT + DGW + Adv | 71.8 | 74.2       | 60.8    | 63.8   |
| DS + DGW  | 71.5    | 73.5       | 61.1    | 64.4   |
| DS + DGW + Adv - L_{sta} | 72.7 | 73.7       | 61.0    | 64.2   |
| DS + DGW + Adv | 73.4 | 74.9       | 62.3    | 66.2   |

TABLE II
EVALUATION ON DIFFERENT PERTURBATION METHODS AND DIFFERENT CONFIGURATIONS OF DGW ON PASCAL VOC2012. THE EXPERIMENTAL RESULTS ARE BASED ON DEEPLAB-V2 BACKBONE. IN THIS EXPERIMENT, OUR ADVERSARIAL DUAL-STUDENT FRAMEWORK IS USED AND WE ONLY MODIFY THE AUGMENTATION STRATEGY $T$.

| Methods      | 1/8 | 1/4 |
|--------------|-----|-----|
| CutMix       | 69.1| 70.7|
| CowMix       | 71.5| 72.4|
| ClassMix     | 72.5| 74.1|
| Basic Affine Transform | 72.3 | 73.9 |
| Rand_{3,3}   | 72.3| 73.9|
| Grid_{3,3} (Rand_{D}) | 72.9 | 74.3 |
| Grid_{3,3} (Rand_{D} + Rand_{D}, ours) | 73.4 | 74.9 |
| Grid_{2,2} (Rand_{S} + Rand_{D}, ours) | 72.8 |   |
| Grid_{4,4} (Rand_{S} + Rand_{D}, ours) | 74.2 |   |
| Grid_{5,5} (Rand_{S} + Rand_{D}, ours) | 72.6 |   |

71.5% and 73.5% for PASCAL VOC2012, from 60.1% and 63.2% to 61.1% and 64.4% for Cityscapes, when 1/8 and 1/4 of labeled data are used, respectively. It verifies that adopting Dual-Student is more robust than Mean-Teacher. From the last two rows, it can be concluded that stabilization loss also remarkably contributes to the final performance.

2) DGW Vs. Other Perturbation Methods: Though ICT [50], CutOut and CowOut [12] made early attempts on unsupervised augmentation for semi-supervised semantic segmentation, we choose to compare DGW with several recent popular unsupervised data augmentation strategies, such as CutMix and CowMix proposed in [11] and ClassMix [13], since they are verified to be better perturbations for the semi-supervised segmentation task. We also compare DGW with basic affine transform which is constructed by random translations, scalings, rotations, and shears. From the top four rows of Table II, we can see that when the overall framework is the same and we replace DGW with other augmentation mechanisms, the performance will be significantly degraded. Specifically, the performance of DGW not only outperforms the class-level augmentation methods, such as Class-Mix, Cut-Mix, and CowMix, but also outperforms the simple geometric-transform-based augmentation, i.e., basic affine transform. The main reason is that the proposed DGW can flexibly enrich the diversity of objects and is image context friendly.

3) Configuration of DGW: Finally, we evaluate how a specific configuration of DGW makes a difference to the performance. In Table II, Rand_{3,3} means the nine source and destination control points are randomly chosen, Grid_{3,3} means to choose the centers of 9 equally divided image grids, Rand_{S} and Rand_{D} denote source and destination points are allowed to have a random Gaussian offsets to the grid centers, respectively. The results show that totally random sampling of control points is not a good choice. It is reasonable because it may result in sever distortion with certain probability. Grid sampling is much better, especially when both source and destination points are allowed to exhibit slight randomness. For block number $n$, we use $n=3$ as the default setting. In Table II, we verify the influence of $n$ and give the corresponding results when $n=2,3,4,5$ respectively. The results show that $n=3$ leads to better results than $n=2$ or 5. Note that, $n=4$ leads to the best result, which demonstrates the potential of our DGW augmentation.

4) Impact of Hyperparameters: There are several hyperparameters in our framework. For the weighting $\lambda's$, we do not tune with them and simply follow previous wisdom in [15] and [47]. Carefully tuning these parameters could bring more impressive results and we leave it as our future work. The settings of $\sigma_s$ and $\sigma_d$ are not quite sensitive to the choice of $\sigma_s$ and $\sigma_d$. We provide empirical results on PASCAL VOC2012 (Table III) for evidence. We adjust $\sigma_s$ and $\sigma_d$ from 0.05 to 0.15 and 0.15 to 0.25, respectively. Results on PASCAL VOC2012 with only 1/8 training data are shown in Table III. From these results, we can see that under a wide range of settings of $\sigma_s$ and $\sigma_d$, our ADS framework can achieve very good mIoU performance on PASCAL VOC2012. When only 1/8 training data is used, the performance is consistently around 73% in terms of mIoU. It shows that our DGW is robust and beneficial. In our experiments, $\sigma_s$ and $\sigma_d$ are set to be 0.1 and 0.2, respectively.

5) Effect of Consistency in Geometric Transformation: To verify the effect of consistency in the geometric transformation, we calculate mIoU between DGW(Net(Input)) and Net(DGW(input)), and the experimental results are shown in Table IV. The higher mIoU means the output of DGW(Network(Input)) is closer to the output of Net(Network(input)), and that is, the network has better geometric transformation consistency. From Table IV, we can see that our proposed method has better geometric transformation consistency than the baseline for each data condition due to our proposed consistency loss. Besides, the model trained on more labeled data not only has a better segmentation result but also has a better geometric transformation consistency than that trained on less labeled data. It verifies that a better model has better geometric transformation consistency, and the consistency loss based on geometric transformation could
improve the performance of segmentation. The qualitative results are shown in Fig. 4, which also demonstrates that our proposed method leads to better geometric transformation consistency and better segmentation results.

### C. Comparison With State-of-the-Arts

1) Results on PASCAL VOC2012: Following previous works, we conduct experiments on eight proportions of labeled data and results are summarized in Table V. Our results are compared with seven recent state-of-the-art works, all of which use the same DeepLab-v2 network. Our method outperforms the supervised baselines, i.e. DeepLab-v2, by a large margin. Besides, the less labeled data, the higher relative improvement. Specifically, when only 1/100 of labeled data is available, our solution improves relatively over baseline by up to 33.3% and with 1/8 training data, we can achieve the same performance of 73.4% in mIoU as fully-supervised DeepLab-v2. These results also demonstrate that our proposed method achieves the best results on PASCAL VOC2012, suggesting that our proposed geometry-based consistency regularization and Adversarial Dual-Student framework can both be successfully applied to semi-supervised semantic segmentation. One reason for the higher performance is that our proposed DGW mechanism introduces complex and flexible geometric perturbation. It enriches diversities in consistency regularization and meanwhile avoids severe context destruction. The other one is that Adversarial Dual-Student can avoid the limitation of MeanTeacher framework and better reveal the potential of unlabeled samples.

The qualitative comparison results are demonstrated in Fig. 5. More visual comparison results are provided in the supplementary file. These is a clear improvement over the baseline and other state-of-the-art methods. Notice that one significant advantage of our method is that the segmentation results are more complete and have less wrong labels.

Besides DeepLab-v2 as the baseline network, we also conduct experiments with DeepLab-v3+ as the baseline. The results are summarized in Table VIII. Note that, the
Fig. 5. Qualitative comparison results on PASCAL VOC2012 using 5% labeled samples. Our proposed semi-supervised approach produces improved results compared to the baseline, and also outperforms other state-of-the-art methods.

TABLE VI

| Methods          | 1/30 | 1/2 | 1/4 | 1/2   | full |
|------------------|------|-----|-----|-------|------|
| DeepLab-v2 [3]   | 43.4 | 55.0| 60.3| 62.2  | 66.2 |
| AdvSemiSeg [27]  | -    | 58.8| 62.3| 65.7  | 67.7 |
| CowMix [12]      | 49.0 | 60.5| 64.1| 66.5  | 69.0 |
| MLMT+GAN [47]    | -    | 59.3| 61.9| -     | -    |
| DST-CBC [28]     | 48.7 | 60.5| 64.4| -     | -    |
| ClassMix [13]    | 54.1 | 61.4| 63.6| 66.3  | -    |
| RCS [31]         | -    | 60.3| 63.8| -     | -    |
| C³-SemiSeg [36]  | 55.2 | 63.2| 65.5| -     | -    |
| Ours             | 56.3 | 62.3| 66.2| 67.4  | 69.0 |
| Rela. Imp (%)    | 29.7 | 13.5| 9.8 | 8.4   | 4.2  |

TABLE VII

Comparing the relative improvements over baselines for different semi-supervised learning methods. When only 1/8 of labeled data is available, our solution improves relatively over baseline by up to 8.6%, which outperforms other state-of-the-art works. When only 1/4 of labeled data is available, our solution achieves the second-best results.
2) Results on Cityscapes and PASCAL-Context: Table VI presents the results on Cityscapes under settings of different labeled data volume. These results demonstrate that our proposed method also significantly and consistently outperforms the baseline model and state-of-the-art semi-supervised methods on Cityscapes. Especially, our proposed method with only 1/4 labeled data achieves the same performance as fully-supervised DeepLab-v2.

The qualitative results with 1/4 labeled samples are demonstrated in Fig 6. Since the differences on the Cityscapes are subtle, in each segmentation map, we add the zoomed-in views of informative areas.

Table VII presents the results on PASCAL-Context dataset with different amounts of labeled data. PASCAL-Context dataset is smaller and more difficult than PASCAL VOC. For this dataset, our proposed method still achieves better performance than the baseline model and state-of-the-art semi-supervised methods. Specifically, when 1/8, 1/4, and full labeled data are used, our solution improves relatively over baseline by 12.5%, 10.5%, and 3.7% respectively.

V. CONCLUSION

In this work, a novel ADS framework with DGW is proposed specially for semi-supervised semantic segmentation task. The proposed DGW is a context-friendly unsupervised data augmentation strategy for unlabeled image perturbation and it is verified to be more effective than existing mechanisms to construct consistency regularization. The ADS framework is also validated to be better than the popular Mean-Teacher framework and state-of-the-art performance can be achieved on both PASCAL VOC2012 and Cityscapes datasets.

Recently, self-paced learning (SPL) [62], [63], [64], [65] is utilized to better exploit the reliable and unreliable labels from the pseudo labels. Among them, adversarial-paced learning (APL) [65] uses adversarial learning to improve the performance and flexibility of self-paced learning. In the future, we would like to improve our adversarial learning based on SPL.

REFERENCES

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” 2014, arXiv:1412.7062.

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput.-Assist. Intervent. Berlin, Germany: Springer, 2015, pp. 234–241.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2017.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 2881–2890.

[6] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, “OCNet: Object context network for scene parsing,” 2018, arXiv:1809.00916.

[7] K. Sun et al., “High-resolution representations for labeling pixels and regions,” 2019, arXiv:1904.04514.

[8] J. Ji, R. Shi, S. Li, P. Chen, and Q. Miao, “Encoder-decoder with cascaded CRFs for semantic segmentation,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 5, pp. 1926–1938, May 2020.

[9] T. Athanasiadis, P. Mylonas, Y. Avrithis, and S. Kollias, “Semantic image segmentation and object labeling,” IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 3, pp. 298–312, Mar. 2007.

[10] F. Meng, K. Luo, H. Li, Q. Wu, and X. Xu, “Weakly supervised semantic segmentation by a class-level multiple group cosegmentation and foreground fusion strategy,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 12, pp. 4823–4836, Dec. 2019.

[11] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, “CutMix: Regularization strategy to train strong classifiers with localizable features,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2019, pp. 6023–6032.

[12] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, “Semi-supervised semantic segmentation needs strong, varied perturbations,” 2019, arXiv:1906.01916.
[58] M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3213–3223.

[59] R. Mottaghi et al., “The role of context for object detection and semantic segmentation in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 891–898.

[60] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.

[61] T.-Y. Lin et al., “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2014, pp. 740–755.

[62] M. P. Kumar, B. Packer, and D. Koller, “Self-paced learning for latent variable models,” in Proc. Adv. Neural Inf. Process. Syst. (Nips), 2010, pp. 1189–1197.

[63] D. Zhang, J. Han, L. Yang, and D. Xu, “SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 475–489, Feb. 2018.

[64] D. Zhang, J. Han, L. Zhao, and D. Meng, “Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework,” Int. J. Comput. Vis., vol. 127, no. 4, pp. 363–380, 2019.

[65] D. Zhang, H. Tian, and J. Han, “Few-cost salient object detection with adversarial-paced learning,” in Proc. NIPS, vol. 33, 2020, pp. 12236–12247.

Cong Cao received the B.E. degree from the China University of Mining and Technology, Xuzhou, China, in 2018, and the M.E. degree from Tianjin University, Tianjin, China, in 2021. After that, he was a Researcher with SenseTime, from March 2021 to June 2022. His research interests include computer vision, computational imaging, and machine learning.

Tianwei Lin received the B.Eng. degree from the School of Mechanical Engineering, Shanghai Jiao Tong University, in 2016, and the master’s degree from Shanghai Jiao Tong University, in 2019. He is a Senior Research and Development Engineer at the Department of Computer Vision (VIS) Technology, Baidu Inc. His research interests include computer vision, deep learning, action recognition, temporal action detection, and GAN.

Dongliang He received the bachelor’s and Ph.D. degrees in computer science from the University of Science and Technology of China, in 2012 and 2017, respectively. He is a Senior Research and Development Engineer at the Department of Computer Vision (VIS) Technology, Baidu Inc. His research interests focus on computer vision, deep learning, and multi-media processing.

Fu Li received the M.E. degree from the Dalian University of Science and Technology, China, in 2015. He is a Senior Research and Development Engineer at the Department of Computer Vision Technology (VIS), Baidu Inc. His research interests focus on video understanding and generative adversarial networks.

Huanjing Yue (Member, IEEE) received the B.S. and Ph.D. degrees from Tianjin University, Tianjin, China, in 2010 and 2015, respectively. She was an Intern with Microsoft Research Asia from 2011 to 2012, and from 2013 to 2015. She has visited the Video Processing Laboratory, University of California at San Diego, from 2016 to 2017. She is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. Her current research interests include image processing and computer vision. She received the Microsoft Research Asia Fellowship Honor in 2013. She was selected into the Elite Scholar Program of Tianjin University in 2017.

Jingyu Yang (Senior Member, IEEE) received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2003, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2009. He has been a Faculty Member with Tianjin University, Tianjin, China, since 2009, where he is currently a Professor with the School of Electrical and Information Engineering. He was with Microsoft Research Asia (MSRA), Beijing, in 2011. He was also with the MSRAs Young Scholar Supporting Program in 2012 and with the Signal Processing Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, from 2014 to 2015. He has authored or coauthored over 140 high quality research papers (including dozens of IEEE Transactions and top conference papers). His research interests include computer vision, computational photography, and 3D vision.

Dr. Yang got the Best 10% Paper Award in IEEE VCIP 2016 and the Platinum Best Paper Award in IEEE ICME 2017, as a coauthor. He served as the Special Session Chair for the International Conference on Visual Communications and Image Processing in 2016 and the Area Chair for the International Conference on Image Processing in 2017. He was selected in the program for New Century Excellent Talents in University (NCET) from the Ministry of Education, China, in 2011, the Tianjin Municipal Innovation Talent Promotion Program in 2015, and a National High-Level Youth Talent Program, China, in 2020.

Errui Ding received the Ph.D. degree from Xidian University in 2008. He is currently the Director of the Computer Vision Technology Department (VIS) and the Augmented Reality Department (AR), Baidu Inc. In recent years, he has published tens of papers on top-tier conferences. He was awarded the Best Paper Runner-Up by ICDAR 2019. He also coorganized two competition tracks, i.e., ART and CSVT, in ICDAR 2019 and the 2nd Workshop on Learning from Imperfect Data in CVPR 2020. He serves as a member for the Special Committee of China Society of Image and Graphics.