Improving Semi-Supervised Semantic Segmentation with Sliced-Wasserstein Feature Alignment and Uniformity

Chen-Yi Lu Kasra Derakhshandeh Somali Chaterji
Purdue University
lu842@purdue.edu, kderakhs@purdue.edu, schaterji@purdue.edu

Abstract

Semi-supervised semantic segmentation with consistency regularization capitalizes on unlabeled images to enhance the accuracy of pixel-level segmentation. Current consistency learning methods primarily rely on the consistency loss between pseudo-labels and unlabeled images, neglecting the information within the feature representations of the backbone encoder. Preserving maximum information in feature embeddings requires achieving the alignment and uniformity objectives, as widely studied. To address this, we present SWSEG*, a semi-supervised semantic segmentation algorithm that optimizes alignment and uniformity using the Sliced-Wasserstein Distance (SWD), and rigorously and empirically proves this connection. We further resolve the computational issues associated with conventional Monte Carlo-based SWD by implementing a Gaussian-approximated variant, which not only maintains the alignment and uniformity objectives but also improves training efficiency. We evaluate* SWSEG *on the PASCAL VOC 2012, Cityscapes, and ADE20K datasets, outshining supervised baselines in mIoU by up to 11.8%, 8.9%, and 8.2%, respectively, given an equivalent number of labeled samples. Further,* SWSEG *surpasses state-of-the-art methods in multiple settings across these three datasets. Our extensive ablation studies confirm the optimization of the uniformity and alignment objectives of the feature representations.*

1. Introduction

Semantic segmentation, the task of assigning pixel-level labels to images, is a cornerstone in domains such as autonomous driving and medical image analysis. Semantic segmentation models typically comprise an encoder and a decoder. The encoder extracts meaningful features from the input images, while the decoder assigns the feature embeddings to desired pixel-level labels. However, its effectiveness is often hampered by the need for costly pixel-level annotated data [9]. To mitigate this, semi-supervised learning (SSL) was proposed [3]. SSL leverages a limited amount of labeled data in conjunction with a larger pool of readily available unlabeled data. The key lies in effectively harnessing the unlabeled data to improve the model's performance.

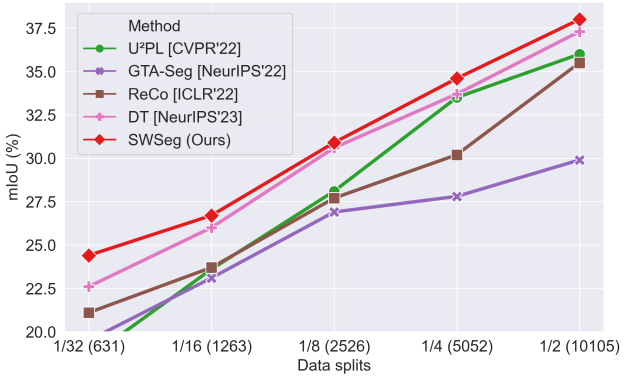


Figure 1. SWSEG demonstrates marked superiority in mIoU across varying data splits of the ADE20K dataset.

Pseudo-labeling and consistency regularization are two prominent techniques in SSL. Pseudo-labeling generates temporary labels from the model's predictions during training and using these to train the model in a supervised manner [2, 26]. The model then computes the loss between these pseudo-labels and the corresponding unlabeled images. This technique allows the model to leverage unlabeled data, thereby improving its performance. Consistency regularization, on the other hand, encourages the model to produce consistent predictions for different perturbations of the same input image, thereby promoting robustness and generalization [14, 32]. This technique is grounded in the smoothness assumption, which posits that data points with similar features should have similar labels [3]. In practice, random augmentations are applied to input images, and the model is trained to minimize discrepancies between its predictions on the original and augmented versions. In the realm of semantic segmentation, consistency regularization is often paired with pseudo-labeling, where weakly augmented unlabeled

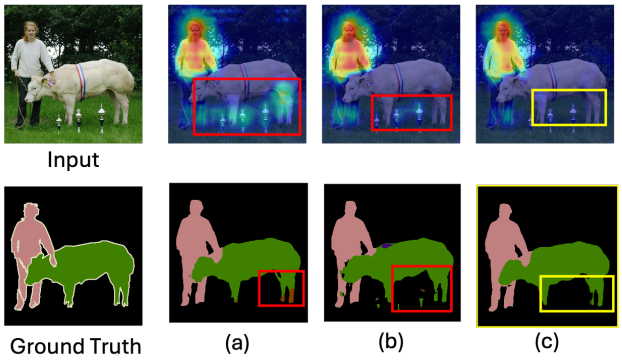


Figure 2. Grad-CAM visualization [33] and model predictions. (a) depicts results from a model trained without alignment optimization, leading to inadequate boundary definition. (b) shows moderate improvements using traditional cross-entropy loss, which partially improves boundary adherence. (c) illustrates SWSEG with SWD loss, significantly enhancing segmentation accuracy by ensuring uniformity and alignment of features.

data generate temporary "pseudo" labels for training [2, 34]. Concurrently, a parallel training stream employs strongly augmented data to calculate and minimize consistency losses, further enhancing model robustness [8, 23, 29, 46, 50].

While consistency-based methods are effective, existing approaches focus solely on enforcing regularization at the decoder's output, neglecting the rich information embedded within the intermediate feature representations from the backbone encoder. This oversight neglects the potential of extending consistency regularization to the feature level, guided by the smoothness assumption. We argue that enforcing consistency within these feature representations, derived from CNN-based or Transformer-based backbone encoders [4, 43], is crucial for robust learning against image perturbations. In Fig. 2, we insert a cross entropy loss on the output of encoders between feature embeddings of of weakly and strongly augmented images. The results shows that it already improves the feature separability and segmentation performance.

To explore the importance of superior feature representations, we draw parallels with the self-supervised contrastive learning literature [1, 6, 48]. These methods, which involve contrasting positive (similar) and negative (dissimilar) pairs, excel at promoting both alignment and uniformity [13, 40]. Alignment, which aligns with the smoothness assumption, posits that perturbations to the same input should yield proximate features. Uniformity, on the other hand, implies that feature representations should be maximally informative, a concept often overlooked in semi-supervised semantic segmentation. These metrics provide a comprehensive framework for evaluating and enhancing feature representations, ultimately leading to improved model performance. The attraction between positive pairs in contrastive learning fosters alignment, while the repulsion of negative pairs encourages diverse and informative representations, enhancing uniformity. Inspired by the effectiveness of contrastive learning, we aim to bridge the gap in semi-supervised semantic segmentation by developing a method that explicitly optimizes both alignment and uniformity of feature representations. This raises a key question: *How can we design an objective function that simultaneously achieves both uniformity and alignment of feature representations?*

In this work, we introduce SWSEG, a semi-supervised semantic segmentation algorithm that directly addresses this question. We leverage the Sliced-Wasserstein distance (SWD), a metric that quantifies the distance between probability distributions, to simultaneously optimize alignment and uniformity of feature representations extracted from the backbone encoder model. SWD achieves this by projecting features onto a unit hypersphere and computing the Wasserstein distance between the resulting one-dimensional embeddings. Empirically, we demonstrate that minimizing SWD effectively optimizes both alignment and uniformity between feature representations. Furthermore, to address the computational burden associated with traditional Monte Carlo estimation of SWD for high-dimensional features, we introduce a variant that projects feature embeddings onto a Gaussian distribution. This not only maintains the uniformity objective, as a normalized Gaussian distribution is uniformly distributed on the unit hypersphere, but also enables efficient computation by leveraging the analytical solution for the quadratic Wasserstein distance between two Gaussian distributions. Additionally, we incorporate a regularization term that decorrelates feature representations, fulfilling the necessary condition for projecting features onto a Gaussian distribution. We evaluated our algorithm on the PASCAL VOC 2012 [12], Cityscapes [9], and ADE20K [51] datasets under various partitioning protocols. Our results consistently demonstrate SWSEG's superiority over supervised baselines, with performance improvements of up to 11.8%, 8.9%, and 8.2% on PASCAL VOC, CityScapes, and ADE20K, using the same amount of labeled data. Notably, SWSEG also achieves state-of-the-art (SOTA) results compared to existing methods under multiple data splitting protocols, as illustrated in Figure 1 [19, 23, 27]. Through extensive ablation studies, we confirm the critical role of uniformity in feature space for effective semi-supervised semantic segmentation. To the best of our knowledge, SWSEG is the first to establish a direct link between SSL and the uniformity metric, while also implementing the SWD to enhance feature representations. Our contributions can be summarized as follows:

1. We introduce SWSEG, a semi-supervised semantic segmentation algorithm that explicitly optimizes feature representation uniformity and alignment using SWD, and back up our claim with extensive empirical studies.

2. We introduce an efficient variant of SWD estimation that projects feature embeddings onto a Gaussian distribution. This approach maintains uniformity while leveraging the analytical solution for quadratic Wasserstein distance between Gaussians, reducing computational complexity.
3. SWSEG achieves SOTA results on PASCAL VOC 2012, Cityscapes, and ADE20K datasets, outperforming supervised baselines and existing semi-supervised methods, with significant performance improvements of up to 11.8% on PASCAL VOC, 8.9% on CityScapes, and 8.2% on ADE20K compared to supervised methods.

2. Related Works

2.1. Semi-Supervised Learning

Semi-supervised learning (SSL) has emerged as a promising solution to alleviate the burden of collecting labeled data, enabling the training of DNNs using only a fraction of labeled data while still achieving impressive results. Pseudo-labeling is a widely adopted technique which leverages the model's predictions on unlabeled data as surrogate labels [21]. This approach has been shown to improve model generalization by shifting the decision boundary towards low-density regions [3]. Additionally, many SSL algorithms incorporate consistency regularization, which regularizes the model to generate consistent outputs regardless of perturbations applied to the input data [20, 35, 38]. Laine *et al*. [20] combined cross-entropy loss with consistency loss for training, and incorporated model predictions into the pseudo-label set through an exponential moving average. Tarvainen and Valpola [38] posited that directly updating the model weights at each mini-batch accelerates the learning process while enhancing the model's performance. Sohn *et al.* [35] generated pseudo-labels from model predictions on weakly-augmented input images and computed the unsupervised loss between these pseudo-labels and the model predictions from strongly-augmented input images. However, these SSL methods were designed for image classification, which is not suitable for semantic segmentation where pixel-wise predictions are desired.

2.2. Semi-Supervised Semantic Segmentation

A number of research papers have incorporated consistency training and pseudo-labeling in SSL for segmentation tasks [29, 37, 39, 42, 44–46, 50]. For instance, Wang *et al.* [42] stored pseudo-labels with lower confidence as negative samples and computed contrastive loss using these samples. Yang *et al.* [46] introduced an additional teacher model and used dropout as feature-wise perturbation for consistency-based pseudo-labeling. However, these methods typically enforce consistency regularization solely on the decoder output, neglecting the rich information present in the intermediate feature representations produced by the backbone encoder. Recent works leveraging feature embedding in SSL segmentation include AllSpark [39], DDFP [41], and CorrMatch [36]. AllSpark balances the importance of labeled and unlabeled training streams by implementing a cross-attention mechanism in the feature embedding. DDFP injects perturbation in the feature space to regularize the decision boundary, with perturbation locations guided by a lightweight normalizing flow-based density estimator. CorrMatch utilizes the correlation map of features from strongly and weakly augmented inputs to regularize pseudo-label supervision in the decoder output. In contrast, SWSEG aims to fill this gap by explicitly focusing on training the backbone encoder to produce more meaningful feature representations for the segmentation head. By optimizing both the alignment and uniformity of these feature representations, SWSEG aims to improve overall segmentation performance. This approach goes beyond the limitations of existing methods that primarily focus on designing complex strategies such as cross-attention layers, normalizing flows, or heuristic label and region propagation techniques.

2.3. Uniformity of Feature Representation

Although existing SSL methods for semantic segmentation have demonstrated promising results through consistency learning, they typically impose consistency on the model's output. While few studies have explored consistency regularization on feature representations from the backbone model, this area remains under-explored compared to extensive research in self-supervised learning [1, 6, 48]. In self-supervised learning, where labeled data is scarce or absent, preventing representational collapse is paramount. This collapse occurs when the model produces uninformative feature representations, regardless of variations in the input. While the uniformity metric, which assesses how evenly features are distributed across the hypersphere, has been effectively used in self-supervised learning to mitigate collapse [40], its application in semi-supervised semantic segmentation has been absent. Recognizing the potential benefits of uniformity in preventing representational collapse and maximizing information preservation, we incorporate this metric into our semi-supervised approach. We further establish a connection between uniformity and the sliced-SWD metric, demonstrating that minimizing SWD between feature representations of differently augmented inputs can simultaneously maximize feature uniformity. This unified approach enhances model performance by ensuring that feature representations are both consistent and informative.

3. Method

This section begins by introducing the conventional framework for consistency learning-based SSL. We then argue for the importance of enforcing consistency regularization in the feature embedding, discussing the roles of alignment and

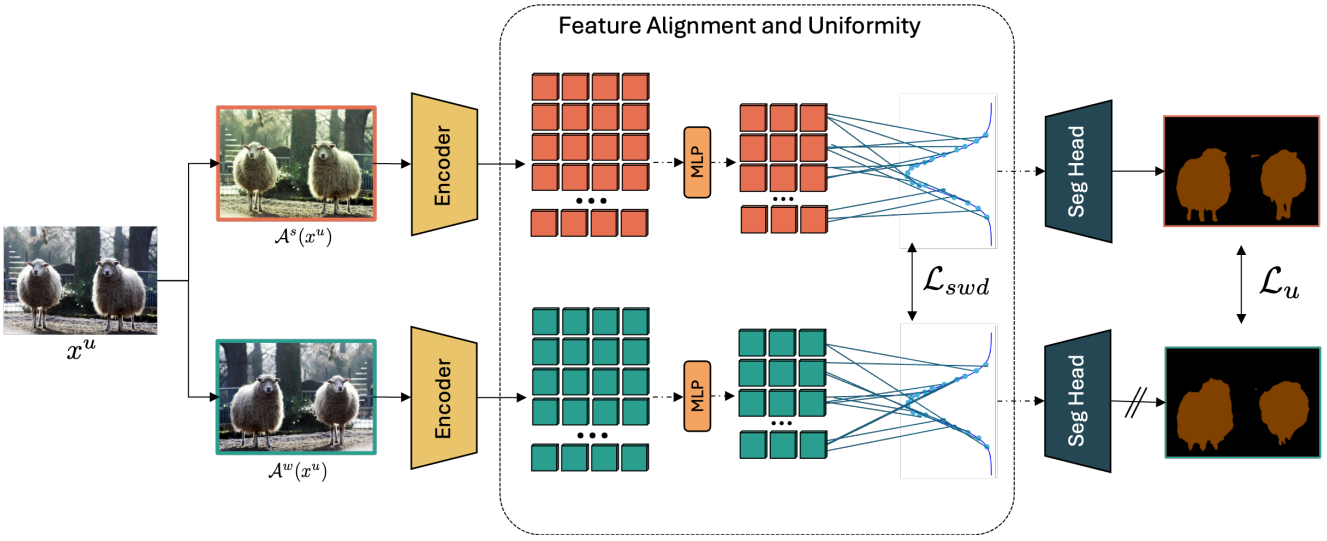


Figure 3. SWSEG Framework for Joint Optimization of Alignment and Uniformity. The SWSEG framework employs both weak ($\mathcal{A}^w$) and strong perturbations ($\mathcal{A}^s$) applied to unlabeled input images x^u. Both augmentations are passed through a shared backbone encoder to extract features, which are then mapped to a different space via non-linear MLPs, producing feature embeddings. To achieve both feature alignment and uniformity, SWSEG introduces an SWD loss $\mathcal{L}_{swd}$, computed between these intermediate feature embeddings. The goal is to encourage the feature embeddings from different augmentations to be consistent (alignment) and to be spread out uniformly in the feature space, maximizing information (uniformity). An unsupervised loss $\mathcal{L}_u$ is computed to ensure consistency between segmentations generated from weakly and strongly augmented inputs.

uniformity in achieving better feature representations. Next, we introduce a Gaussian-projected SWD metric to simultaneously achieve uniformity and alignment in the feature space. Finally, we detail the integration of these elements within the SWSEG framework.

3.1. Preliminary

In semi-supervised learning, we are given labeled image sets, $D^l = \{x_i^l, y_i^l\}$, and unlabeled ones, $D^u = \{x_j^u\}$. Typically, D^u greatly outnumbers D^l. Our goal is to leverage these unlabeled images to optimize the parameters θ of a semantic segmentation model $f_\theta(\cdot)$.

In SWSEG, we adopt a typical framework from FixMatch [34], training the model through two separate feed-forward streams: supervised and unsupervised. In the supervised stream, we draw a batch of images from D^l and compute the cross-entropy loss $\mathcal{L}_s$ between the model's predictions $\hat{y}_i^l$ and the labels y_i^l. Conversely, the unsupervised stream draws images from D^u, to which we apply weak or strong perturbations, $\mathcal{A}^w(\cdot)$ and $\mathcal{A}^s(\cdot)$, respectively. The unsupervised loss $\mathcal{L}_u$ is then minimized according to the equation (1):

$$\mathcal{L}_u = \frac{1}{|\mathcal{D}^u|} \sum_{x \epsilon \mathcal{D}^u} \mathcal{H}\left(f_\theta\left(\mathcal{A}^s\left(x^u\right)\right), f_\theta\left(\mathcal{A}^w\left(x^u\right)\right)\right), \quad (1)$$

We generate pseudo-labels $f_\theta(\mathcal{A}^w(x^u))$ by feeding weakly-augmented images into the model. We then compute the unsupervised loss using strongly-augmented samples $f_\theta(\mathcal{A}^s(x^u))$, using $\mathcal{H}$, for cross-entropy loss. The pseudo-label generation does not contribute to gradient computation.

3.2. Feature Alignment and Uniformity

In semantic segmentation, $f_\theta(\cdot)$ typically comprises a backbone encoder p_θ and a segmentation head q_θ [4, 43]. Existing methods [23, 27, 37, 46, 50] focus on enforcing consistency regularization on the segmentation head's output, often overlooking backbone encoder features. To further harness the potential of feature representations, we draw on self-supervised contrastive learning, which trains networks to generate meaningful embeddings without labels, emphasizing alignment and uniformity as key objectives for assessing feature quality.

Alignment aims to minimize differences between embeddings from similar samples. Intuitively, model should produce similar feature embeddings when images with same classes are input. We achieve this by varying the magnitude of perturbations applied to the input images and minimizing the distance between embeddings using a distance metric dist:

$$\mathcal{L}_{align} = \text{dist}\left(p_\theta\left(\mathcal{A}^s\left(x_i^u\right)\right), p_\theta\left(\mathcal{A}^w\left(x_i^u\right)\right)\right). \quad (2)$$

Uniformity promotes even distributions of feature representations across a hypersphere, enhancing model robustness by preventing collapse to a lower-dimensional subspace.

While often less emphasized in semi-supervised learning due to the stabilizing influence of labeled data, uniformity is critical in self-supervised learning to prevent the model from generating identical embeddings regardless of input. By ensuring uniformity, the model preserves a rich diversity of information within the feature space, which significantly improves the performance of the segmentation head.

3.3. Revisiting Sliced Wasserstein Distance

Intuitively, the uniform distribution on a unit hypersphere exemplifies maximum uniformity. We need to establish an objective function that can achieve both uniformity and alignment. To this end, we implement the SWD as our distance metric. SWD calculates average distances from 1-D projections sampled from a unit hypersphere, thus inherently promoting both desired properties.

The rationale behind SWD stems from its ability to reduce the computational complexity relative to the high-dimensional Wasserstein distances (WD). Unlike the traditional WD, which struggles with high dimensionality, SWD leverages the 1-dimensional Wasserstein distance, which can be computed in closed form: Let μ_d and ν_d be two probability measures in $\mathcal{P}(\Omega) \in \mathbb{R}^d$:

$$W_p^p(\mu_d, \nu_d) = \left(\int_0^1 |F_{\mu_d}^{-1}(z) - F_{\nu_d}^{-1}(z)|^p \, dz \right)^{1/p}, \quad (3)$$

where p is the order of the distance, and $F_{\mu_d}^{-1}$ and $F_{\nu_d}^{-1}$ are the quantile functions of μ_d and ν_d, respectively. Computationally, this is achieved by sorting and computing the mean l^2-distance between two 1-D vectors. SWD [30] involves projecting high-dimensional features onto a unit hypersphere and calculating the 1-D Wasserstein distance between these projections. Formally, let $\mathbb{S}^{d-1}$ represent the unit sphere in d dimensions, σ the uniform distribution on $\mathbb{S}^{d-1}$, and $P_{\theta\#}\xi$ the push-forward measure of $\xi \in \mu_d, \nu_d$. The SWD is defined as:

$$SW_p^p(\mu_d, \nu_d) = \int_{\mathbb{S}^{d-1}} W_p^p(P_{\theta\#}\mu_d, P_{\theta\#}\nu_d) \, \mathrm{d}\sigma(\theta). \quad (4)$$

To approximate SWD, we vectorize the input distributions μ_d and ν_d from $\mathbb{R}^{n\times n\times d}$ to $\mathbb{R}^{n^2 d}$. We then employ a Monte Carlo scheme, sampling L directions from the unit sphere $\mathbb{S}^{d-1}$. Projecting μ_d and ν_d along these directions yields L one-dimensional projections for each distribution. After sorting these projections, we calculate the 1D Wasserstein distance between them using Eq. (3). This method efficiently estimates the Wasserstein distance between feature embeddings, exploiting the closed form property of 1D calculations compared to high-dimensional alternatives.

Although the unit sphere projection enables uniformity and minimizing 1-D Wasserstein distances facilitates alignment, training with Monte Carlo-approximated SWD presents several challenges. First, the results are inherently non-deterministic and intractable due to the randomness inherent in the Monte Carlo method. Second, as the dimensions of μ_d and ν_d increase, both the computational cost and memory requirements for SWD escalate. Finally, to maintain accuracy in the face of increasing dimensions, the number of directions L must also increase, complicating the Monte Carlo approximation further.

3.4. Proposed Approach: SWSEG

To enhance feature representation uniformity and alignment while alleviating the computational burden of Monte Carlo-based SWD, we implement an alternate method proposed by Nadjahi *et al*. [28] for approximating the SWD. Our method diverges by projecting these measures directly onto a Gaussian distribution. This modification offers significant computational benefits and supports key theoretical objectives in two primary ways: First, projecting onto Gaussian distributions allows the 2-Wasserstein distance between any two Gaussians to be expressed in a closed-form equation. Second, using normalized Gaussian variables ensures that the resulting distributions are uniformly distributed over the unit hypersphere, which we prove in the supplementary materials. This uniform distribution is essential for maintaining the model's uniformity objective, ensuring features are evenly distributed in the embedding space, maximizing its information for segmentation head.

To formalize this idea, for any positive integer d, let $\{\theta_i\}_{i=1}^d$ be a sequence of i.i.d. 1-d standard Gaussian distributions and $\{X_i\}_{i=1}^d$ be a sequence of 1-d probability distributions. Several Central Limit Theorems [11, 31] ensure that the sequence of distributions formed by the inner product $d^{-1/2}\langle\theta_{1:d}, X_{1:d}\rangle$ converges in probability to a Gaussian random variable when the distributions exhibit weak correlation. Furthermore, according to Proposition 2 from [28], as the dimensions of the distributions increase, the quadratic Wasserstein distance between two zero-mean Gaussian distributions with variance $d^{-1}m_2(\bar{\mu_d})$ can be approximated by the SW distance between two Gaussian projected distributions. The Wasserstein distance between two Gaussian probability distribution is in closed form [10], which enables us to approximate the l^2-SW distance by:

$$\mathcal{L}_{swd} = d^{-1}(m_2(\bar{\mu_d})^{\frac{1}{2}} - m_2(\bar{\nu_d})^{\frac{1}{2}}) + d^{-1}|m_{\mu_d} - m_{\nu_d}|_2^2. \quad (5)$$

This equation comprises two terms: the first quantifies the difference in the second moments of the distributions, and the second captures the squared error in their means. For $\xi_d \in \{\bar{\mu_d}, \bar{\nu_d}\}$, we define $m_2(\xi_d) = n^{-1}\sum_{j=1}^n ||x^j||^2$, enabling a closed-form deterministic solution for $\mathcal{L}_{swd}$. Nadjahi *et al.* also proved in Proposition 1 that $\mathcal{L}_{swd}$ with Gaussian projections equals the original SW_p^p when $p = 2$. Thus, we can use $\mathcal{L}_{swd}$ as a computationally efficient Wasserstein metric for our consistency loss term. A PyTorch-styled

Table 1. Results on ADE20k. The fractions represent the labeled image proportions sampled during training. The model is trained on SegFormer [43] with MiT-B1 backbone. We use mIOU as our evaluation metric.

Method	1/32 (631)	1/16 (1263)	1/8 (2526)	1/4 (5052)	1/2 (10105)
Supervised Only	15.2	22.3	25.9	29.5	32.6
CPS [8] [ICLR '21]	18.8	22.3	27.9	32.4	36.9
U^2PL [42] [CVPR '22]	18.5	23.6	28.1	33.5	36.0
ReCo [22] [ICLR '22]	21.1	23.7	27.7	30.2	35.5
GTA-Seg [19] [NeurIPS '22]	19.5	23.1	26.9	27.8	29.9
DT [27] [NeurIPS '23]	22.6	26.0	30.6	33.7	37.3
SWSEG (Ours)	**24.4**	**26.7**	**30.9**	**34.6**	**38.0**

pseudo code is presented in Algorithm 1. Furthermore, to ensure a more accurate approximation of the Gaussian-SWD and comply with the relaxed Central Limit Theorems in [11, 31], minimizing strong correlations among input embeddings is essential. We address this by introducing a regularization term based on the input cross-covariance matrix into the SWD loss:

$$\mathcal{L}_{reg} = |\mathrm{Cov}(X) - \mathrm{Diag}(X)|_2^2 + |\mathrm{Cov}(Y) - \mathrm{Diag}(Y)|_2^2. \tag{6}$$

Here, $\mathrm{Cov}(\cdot)$ calculates the covariance matrix of the input features, and $\mathrm{Diag}(\cdot)$ extracts its diagonal elements. Minimizing the off-diagonal elements decorrelates the features, thereby indirectly promoting the uniformity of the distribution of the features across the representation space [49].

We present a detailed overview of this training scheme in Fig. 3. We also insert an additional non-linear MLP projection layer before loss computation between representations, following recent self-supervised learning work [6, 7, 16]. The projection layer is solely incorporated during the training process, and it is subsequently discarded when the model is deployed for inference. Our objective is to retain more information in the features prior to the MLP layer. This will enable the segmentation head to utilize the preserved information more effectively for the segmentation task. The specific architecture of the projection layer used in this work consists of two $\mathsf{Conv} - \mathsf{BatchNorm} - \mathsf{ReLu}$ blocks followed by a single Conv layer. The overall training objective, managing the balance of various losses, is summarized in Eq. (7). The hyperparameters λ_1, λ_2, and λ_3 control the weight of each loss.

$$\mathcal{L}_{overall} = \mathcal{L}_s + \lambda_1 \mathcal{L}_u + \lambda_2 \mathcal{L}_{swd} + \lambda_3 \mathcal{L}_{reg}. \tag{7}$$

Algorithm 1 Pseudocode of our Gaussian-SWD loss in a PyTorch-like style.

```
def gaussian_swd_loss(x, y):
    # Reshape the input tensors to (n, dim)
    x, y = x.view(x.shape[0], -1), y.view(y.shape
        [0], -1)
    n, dim = x.shape

    meanx = torch.mean(x, dim=0)
    xc = x - meanx # Zero mean
    m2x = torch.mean(torch.linalg.norm(xc, dim=1)
        ** 2) / dim

    meany = torch.mean(y, dim=0)
    yc = y - meany # Zero mean
    m2y = torch.mean(torch.linalg.norm(yc, dim=1)
        ** 2) / dim

    m_t = torch.linalg.norm(meanx - meany) ** 2 /
        dim
    swd = m_t + (m2x ** (1/2) - m2y ** (1/2)) ** 2
    return swd
```

4. Evaluation

4.1. Datasets

ADE20K. [51] is a complex scene parsing dataset featuring 150 objects, with 25,574 training, 2,000 validation, and 3,000 testing images. For a fair comparison, we used the data split from dual-teacher [27].

PASCAL VOC 2012. [12] This dataset is a classic semantic segmentation dataset comprising 20 classes of objects and 1 background class, with 1464 training, 1449 validation, and 1546 testing high-quality images in the standard set. We incorporate the augmented set from [17], following previous works, increasing our labeled data to 10,582 images.

Cityscapes. [9] This is an autonomous driving semantic segmentation dataset that consists of 19 semantic classes, with 2975, 500, and 1525 training, validation, and testing images, respectively.

4.2. Implementation details.

For ADE20K, we implemented our method with SegFormer [43] architecture with MiT-B1 backbone. For PASCAL VOC 2012 and CityScapes, to have a fair comparison with previous benchmark semi-supervised semantic segmentation methods [27, 37, 42, 45, 46, 50], we implement our method using ResNet-101 [18] as the backbone encoder, initialized with ImageNet pre-trained weights, and

Table 2. Results on CityScapes dataset. The fractions represent the proportion of labeled images used during training. We use mIOU as our evaluation metric.

Method	1/16	1/8	1/4
Supervised Only	67.9	73.2	75.1
PCR [44] [NeurIPS '22]	73.4	76.3	78.4
UniMatch [46] [CVPR '23]	76.6	77.9	79.2
AugSeg [50] [CVPR '23]	75.2	77.8	79.6
DT [27] [NeurIPS '23]	76.8	78.4	79.5
DAW [37] [NeurIPS '23]	76.6	78.4	79.8
CorrMatch [36] [CVPR '24]	**77.3**	78.5	79.4
RankMatch [24] [CVPR '24]	77.1	**78.6**	**80.0**
DDFP [41] [CVPR '24]	77.1	78.2	79.9
SWSEG (Ours)	76.8	**78.6**	79.5

Table 3. Results on PASCAL VOC 2012 dataset. The fractions represent the proportion of labeled images used during training. We use mIOU as our evaluation metric.

Method	1/16	1/8	1/4
Supervised Only	67.2	71.3	73.5
UniMatch [46] [CVPR '23]	75.0	76.8	77.5
AugSeg [50] [CVPR '23]	77.0	77.3	78.8
DAW [37] [NeurIPS '23]	78.5	78.9	79.6
CorrMatch [36] [CVPR '24]	78.4	**79.3**	79.6
RankMatch [24] [CVPR '24]	78.9	79.2	**80.0**
DDFP [41] [CVPR '24]	78.3	78.9	79.8
SWSEG (Ours)	**79.0**	**79.3**	79.9

DeeplabV3+ [5] as the decoder. We use SGD optimizer with momentum of 0.9. The learning rate is set to 0.001, 0.02, and 5e-5 for PASCAL, CityScapes, and ADE20K, respectively. We utilize a polynomial scheduling strategy with 0.9 decay rate.

Data augmentations. For weak augmentations, horizontal flipping, random cropping, and random scaling are applied to the training images. For strong augmentations, we apply random color jittering, random grayscale, random blurring, and Cutmix [47] to the input training images. The training image size is set as 513×513, 769×769, and 512×512 for PASCAL VOC, CityScapes, and ADE20K, respectively. The training epochs for VOC, CityScapes, and ADE20K is 80, 240, and 200 epochs.

Evaluation. As in previous works, we employ mean Intersection-over-Union (mIOU) as our evaluation metric. For the PASCAL VOC 2012 and ADE20K, evaluations are on the center-cropped validation set, while for CityScapes, we employ a sliding window evaluation.

Table 4. Ablation analysis of feature alignment loss functions, $\mathcal{L}_{reg}$, and MLP layer. The results reveal that MSE, cross entropy, and Monte Carlo estimated SWD loss with $L = 128$ yield inferior results. The combination of Gaussian-SWD (GSWD) with $\mathcal{L}_{reg}$ and MLP produces the best performance, with an mIoU of 79.0.

Loss Functions				$\mathcal{L}_{reg}$	MLP	mIoU
MSE	CE	SWD	GSWD			
✓				✓	✓	70.2
	✓			✓	✓	77.1
		✓		✓	✓	77.6
			✓			77.4
			✓	✓		78.1
			✓	✓	✓	**79.0**

Table 5. Analysis of the Uniformity value $\mathcal{L}_{\texttt{uniform}}$ across different data splits of the PASCAL VOC dataset demonstrates that training with the SWD loss $\mathcal{L}_{swd}$ yields superior Uniformity values compared to training without this loss term.

Method	1/8	1/2	1
Without $\mathcal{L}_{swd}$	-2.5896	-2.7095	-2.5574
With $\mathcal{L}_{swd}$	**-2.6592**	**-2.7104**	**-2.6168**

4.3. Comparison with State-of-the-Arts

Results on ADE20K Dataset. Tab. 1 demonstrates the performance of SWSEG against the supervised baseline and existing methods. SWSEG significantly improves over the supervised baseline by 8.2%, 3.4%, 3.7%, 5.1%, and 5.4% mIoUs on the 1/32, 1/16, 1/8, 1/4, and 1/2 splits, respectively. Our method also outperforms the state-of-the-art across all data splits, showcasing the generalizability of SWSEG across different model architectures and datasets.

Results on Cityscapes Dataset. Tab. 2 presents our results in comparison to the supervised baseline and SOTA methods. SWSEG outperforms the supervised baseline by 8.9%, 5.3%, and 4.4% for the 1/16, 1/8, and 1/4 partitions, respectively. Additionally, SWSEG surpasses the previous state-of-the-art method, DAW, by 0.2% in the 1/16 and 1/8 partitions, and performs on par with contemporary methods.

Results on PASCAL VOC 2012 Dataset. Tab. 3 compares SWSEG with existing state-of-the-arts on PASCAL VOC 2012 set. We follow the data partitions in CPS [8]. We outperform the supervised baselines by 11.8%, 8.0% and 6.4% mIoUs under 1/16, 1/8, and 1/4 partitioning protocols. Our method also beats the SOTA DDFP [37] by 0.7%, 0.4%, and 0.1% mIoUs under 1/16, 1/8, and 1/4 splits, respectively. We also visualized the predicted features in Fig. 4 to evaluate the effectiveness of our algorithm's clustering capabilities. Our method demonstrates enhanced discriminative clustering compared to the Supervised Only and UniMatch methods.

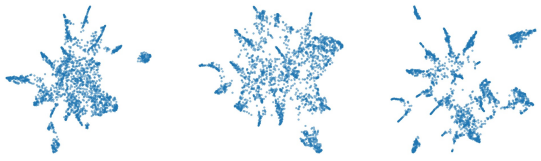
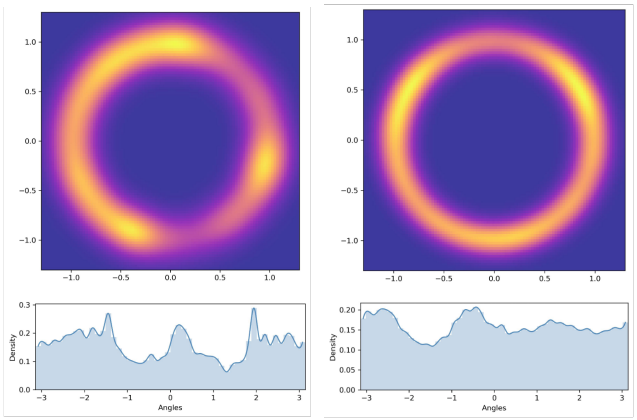


Figure 4. UMAP visualization [25] of feature embeddings using models trained with: (a) Supervised Only, (b) UniMatch, and (c) SWSEG. These visualizations utilize the PASCAL VOC 2012 dataset. Our method demonstrates enhanced clustering capabilities, outperforming both the Supervised Only and UniMatch.



(a) Without $\mathcal{L}_{swd}$ (b) With $\mathcal{L}_{swd}$

Figure 5. We visualize feature representations on a unit circle for models trained with and without $\mathcal{L}_{swd}$ for uniformity analysis. Features from the model with $\mathcal{L}_{swd}$ (b) are more uniformly distributed than those without $\mathcal{L}_{swd}$ (a).

4.4. Ablation Study

Uniformity and Alignment. We conduct an empirical analysis of feature representation alignment and uniformity (Tab. 4). We compare different loss functions for enforcing consistency regularization on the feature representation. Our results indicate that, aside from MSE loss, additional loss terms generally enhance segmentation prediction performance, highlighting the critical role of feature alignment. Monte Carlo SWD with $L = 128$ outperforms MSE and cross-entropy but encounters out-of-memory issues at $L = 1000$ using an identical number of GPUs, highlighting its computational cost. Conversely, our Gaussian-based SWD approximation delivers optimal results with lower memory usage. For uniformity analysis, we reduce the dimension of feature embeddings using PCA [15], projecting them into $\mathbb{R}^2$ and plotting their coordinates on a unit circle, as shown in Fig. 5. The visualization demonstrates that models trained with $\mathcal{L}_{swd}$ achieve greater uniformity compared to those without it, confirming the effectiveness of our introduced SWD loss. To quantitatively validate this observation, we compute the uniformity metric $\mathcal{L}_{\texttt{uniform}}$ from [40, Equation 5.2] and compare the values for models trained with and without $\mathcal{L}_{swd}$ in Tab. 5. The results show that including $\mathcal{L}_{swd}$ decreases the uniformity loss, which aligns with our qualitative visualization in Fig. 5. We also conduct an ablation study on MLP layers and $\mathcal{L}_{reg}$ in Tab. 4, showing that including all components yields the best outcome.

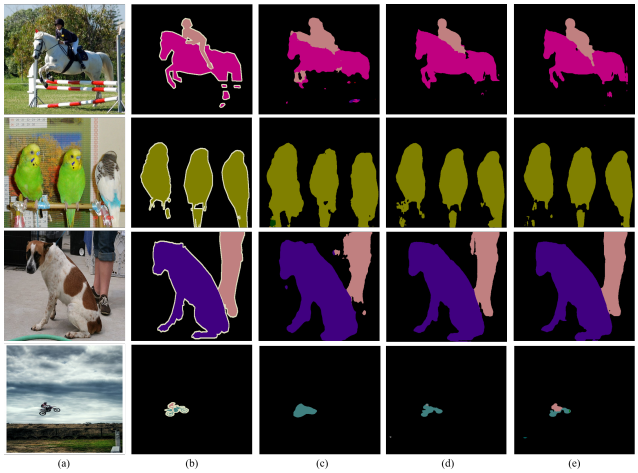


Figure 6. Qualitative results on PASCAL VOC 2012 validation set. All models are trained under 1/4 partition. (a) Input images. (b) Ground Truth labels. (c) Model output when trained with only labeled images. (d) Model output when training without SWD loss. (e) Predictions from SWSEG. When including SWD loss, the predictions are better in ambiguous regions and object borders.

Qualitative analysis. Fig. 6 shows the results of different training methods on the PASCAL VOC 2012 val set. When training with SWD loss, our method outperforms other methods, producing segmentation predictions that are closer to ground-truth labels. We can see from the results that our method performs exceptionally well on the border between different classes. Moreover, the segmentation results of our method have a cleaner prediction between the border and the backgrounds.

5. Conclusion

We introduced SWSEG, a semi-supervised semantic segmentation algorithm designed to enhance both alignment and uniformity of feature representations. We implement the Sliced-Wasserstein Distance (SWD) as an additional loss to jointly optimize these aspects, using a Gaussian-approximated variant to reduce computational overhead while preserving uniformity. Empirical evaluations on PASCAL VOC 2012, CityScapes, and ADE20K datasets show that SWSEG outperforms both supervised baselines and existing methods. Extensive ablation studies highlight the critical role of uniformity in the feature space for effective semi-supervised semantic segmentation. SWSEG is the first method to link semi-supervised learning with the uniformity metric using SWD, marking a significant advancement in the field.

6. Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Numbers CNS-2333487 (NSF Frontier award) and CNS-2146449 (NSF CAREER award) and DEVCOM ARL Army Research Office under Contract number W911NF-2020-221. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. The authors thank the reviewers for their enthusiastic comments and their valuable insights that improved our paper.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 2, 3

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[3] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *International workshop on artificial intelligence and statistics*, pages 57–64. PMLR, 2005. 1, 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2, 4

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 6

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 6

[8] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 2, 6, 7

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1, 2, 6

[10] D.C. Dowson and B.V. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 5

[11] Lutz Dümbgen, Del Conte-Zerial, et al. On low-dimensional projections of high-dimensional distributions. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 91–105. Institute of Mathematical Statistics, 2013. 5, 6

[12] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007:1–45, 2012. 2, 6

[13] Xianghong Fang, Jian Li, Qiang Sun, and Benyou Wang. Rethinking the uniformity metric in self-supervised learning. *arXiv preprint arXiv:2403.00642*, 2024. 2

[14] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 1

[15] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901. 8

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284. Curran Associates, Inc., 2020. 6

[17] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[19] Ying Jin, Jiaqi Wang, and Dahua Lin. Semi-supervised semantic segmentation via gentle teaching assistant. *Advances in Neural Information Processing Systems*, 35:2803–2816, 2022. 2, 6

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[21] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896, 2013. 3

[22] Shikun Liu, Shuaifeng Zhi, Edward Johns, and Andrew J Davison. Bootstrapping semantic segmentation with regional contrast. *arXiv preprint arXiv:2104.04465*, 2021. 6

[23] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022. 2, 4

[24] Huayu Mai, Rui Sun, Tianzhu Zhang, and Feng Wu. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3391–3401, 2024. 7

[25] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. 8

[26] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 1

[27] Jaemin Na, Jung-Woo Ha, Hyung Jin Chang, Dongyoon Han, and Wonjun Hwang. Switching temporary teachers for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2023. 2, 4, 6, 7

[28] Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli. Fast approximation of the sliced-wasserstein distance using concentration of random projections. *Advances in Neural Information Processing Systems*, 34:12411–12424, 2021. 5

[29] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-Supervised Semantic Segmentation With Cross-Consistency Training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12671–12681, Seattle, WA, USA, 2020. IEEE. 2, 3

[30] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012. 5

[31] Galen Reeves. Conditional central limit theorems for gaussian projections. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3045–3049. IEEE, 2017. 5, 6

[32] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 1

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[34] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 2, 4

[35] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 3

[36] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 7

[37] Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: Exploring the better weighting function for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2023. 3, 4, 6, 7

[38] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3

[39] Haonan Wang, Qixiang Zhang, Yi Li, and Xiaomeng Li. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. *arXiv preprint arXiv:2403.01818*, 2023. 3

[40] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 2, 3, 8

[41] Xiaoyang Wang, Huihui Bai, Limin Yu, Yao Zhao, and Jimin Xiao. Towards the uncharted: Density-descending feature perturbation for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3303–3312, 2024. 3, 7

[42] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 3, 6

[43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 2, 4, 6

[44] Haiming Xu, Lingqiao Liu, Qiuchen Bian, and Zhen Yang. Semi-supervised semantic segmentation with prototype-based consistency regularization. *Advances in Neural Information Processing Systems*, 35:26007–26020, 2022. 3, 7

[45] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4268–4277, 2022. 6

[46] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. 2, 3, 4, 6, 7

[47] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 7

[48] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2, 3

[49] Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. Zero-cl: Instance and feature decorrelation for negative-free symmetric contrastive learning. In *International Conference on Learning Representations*, 2021. 6

[50] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11350–11359, 2023. 2, 3, 4, 6, 7

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 6