TOWARDS UNBIASED LEARNING IN SEMI-SUPERVISED SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

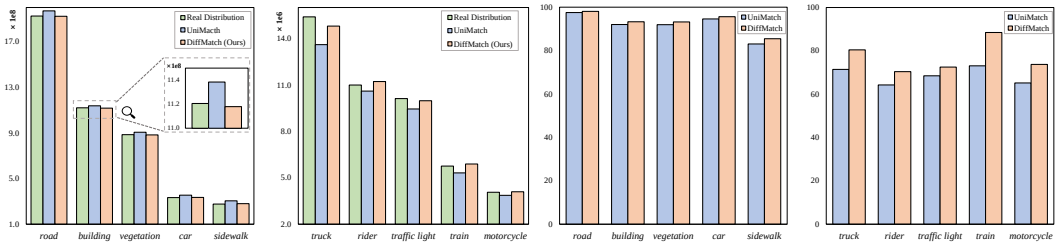
ABSTRACT

Semi-supervised semantic segmentation aims to learn from a limited amount of labeled data and a large volume of unlabeled data, which has witnessed impressive progress with the recent advancement of deep neural networks. However, existing methods tend to neglect the fact of class imbalance issues, leading to the Matthew effect, that is, the poorly calibrated model's predictions can be biased towards the majority classes and away from minority classes with fewer samples. In this work, we analyze the Matthew effect present in previous methods that hinder model learning from a discriminative perspective. In light of this background, we integrate generative models into semi-supervised learning, taking advantage of their better class-imbalance tolerance. To this end, we propose DiffMatch to formulate the semi-supervised semantic segmentation task as a conditional discrete data generation problem to alleviate the Matthew effect of discriminative solutions from a generative perspective. Plus, to further reduce the risk of overfitting to the head classes and to increase coverage of the tail class distribution, we mathematically derive a debiased adjustment to adjust the conditional reverse probability towards unbiased predictions during each sampling step. Extensive experimental results on various domains (natural image/remote sensing image/medical image domains) across multiple benchmarks, especially in the most limited label scenarios with the most serious class imbalance issues, demonstrate that DiffMatch performs favorably against state-of-the-art methods. Code and models will be made available to facilitate future research.

1 INTRODUCTION

Machine learning, especially deep learning, has been consistently reported to achieve competitive or even superior performance compared to human beings in certain supervised learning tasks (LeCun et al., 2015; He et al., 2016a). In real-world scenarios, however, its data-driven nature makes it heavily dependent on massive annotations, especially at the dense pixel level, which is laborious and time-consuming to gather (taking semantic segmentation as a case study). To alleviate the data-hunger issue, considerable works (Wang et al., 2023b; Na et al., 2023; Wang et al., 2023a; Liang et al., 2023) have turned their attention to semi-supervised semantic segmentation in pursuit of bypassing the labeling cost, demonstrating great potential in widespread applications (Siam et al., 2018; Asgari Taghanaki et al., 2021). Since only limited labeled data is accessible, how to fully exploit a large volume of unlabeled data to improve the model's generalization performance for robust segmentation is thus extremely challenging. To leverage unlabeled data effectively, pseudo-labeling (Lee et al., 2013; Rizve et al., 2021) and consistency regularization (Sajjadi et al., 2016; Laine & Aila, 2016) have emerged as mainstream paradigms for semi-supervised segmentation. Recently, these two paradigms are often assembled in the form of a teacher-student scheme (Wang et al., 2022a; Chen et al., 2023a). In this scheme, the teacher network, with a weakly augmented view, generates pseudo labels to guide the counterparts from the student network in the presence of a strongly augmented view, following the smoothness assumption (Chapelle et al., 2009).

From the perspective of probabilistic modeling, almost all de facto methods can be unified as discriminative models, which directly model the conditional probability of discriminating different values across classes for given pixels of an image (*i.e.*, maximizing posterior probability). Despite yielding promising results, these methods tend to neglect the fact of class imbalance issue (*i.e.*, long-tailed distribution). For example, the pixel count of head class *road* can be hundreds of times



(a) **Head Classes / *Pseudo Label Count*** (b) **Tail Classes / *Pseudo Label Count*** (c) **Head Classes / *IoU*** (d) **Tail Classes / *IoU***

Figure 1: We count the training samples of each class on Cityscapes (Cordts et al., 2016) under 1/16 partition protocols and compare the proposed DiffMatch with the highly competitive UniMatch (Yang et al., 2022) in terms of *Pseudo Label Count* and *IoU*, assuming that the ground truth for unlabeled data is available solely for theoretical analysis purposes. (a) Prediction distribution of head classes. (b) Prediction distribution of tail classes. (c) Performance of head classes. (d) Performance of tail classes. Our DiffMatch strives to mitigate the Matthew effect raised by the class imbalance issue and stands out for head/tail classes.

larger than that of tailed class *motorcycle* in the widely used Cityscapes dataset (Cordts et al., 2016) as shown in Figure 1. This highly skewed distribution can lead to the Matthew effect; that is, the poorly calibrated model's predictions can be *biased* towards the majority classes and away from minority classes with fewer samples. This is a corollary raised by discriminative models, which only learn decision boundaries between classes while disregarding the underlying distribution. In other words, these methods, by minimizing empirical risk under the assumption of low-density separation, are highly fragile to the number of pixels per class (*i.e.*, class imbalance), leading to decision boundaries that can be drastically altered by the majority classes (*i.e.*, confirmation bias (Guo et al., 2017)). This affects the quality of pseudo labels, and then aggressively training with erroneous pseudo labels, in turn, exacerbates the model's bias in a self-reinforcing manner, compromising performance. For example, UniMatch (Yang et al., 2022) tends to prioritize the head classes in Figure 1 (a) over tail classes in Figure 1 (b) in terms of pseudo label count compared to real distribution. To make matters worse, the negative impact is inevitably amplified by inbuilt low-data regimes of semi-supervised segmentation, hindering the learning process. Then, the question naturally arises: *How to effectively alleviate the negative impact raised by class imbalance issue and move towards unbiased learning?*

In this work, we analyze the Matthew effect present in previous methods that hinder the model's learning when dealing with class imbalance issues from a *discriminative perspective*. Compared with the discriminative models, the generative models conceptually exhibit better class-imbalance tolerance, attributed to their better asymptotic error approaching rate (Ng & Jordan, 2001) (detailed in Appendix A). In light of this background, we turn to formulate the semi-supervised semantic segmentation task as a conditional discrete data generation problem to model the underlying distribution, alleviating the Matthew effect of discriminative solutions from a *generative perspective*. To this end, we propose DiffMatch to learn a series of state transitions under the guidance of the input image, transforming noise from a known noise distribution into a prediction that better matches the real distribution, maximizing the mutual information between the learned distribution and the underlying real one. A heuristic explanation of the transition process is that it can be viewed as the human process of discriminating objects, gradually scrutinizing them closer after an initial glance with the naked eye. By formulating the pseudo-label generation of the teacher-student scheme as an optimization problem progressively solved by the denoising diffusion process, DiffMatch favors a better capacity to tackle the severe class imbalance issue in semi-supervised learning. Plus, to further reduce the risk of overfitting to the head classes and to increase coverage of the tail class distribution, we mathematically derive a debiased adjustment based on the state transition function of the diffusion process to adjust the conditional reverse probability towards unbiased predictions during each sampling step. This adjustment, formalized as an additional regularization term, further unlocks the potential of DiffMatch to mitigate the Matthew effect effectively and is in line with the step-by-step sampling nature of the diffusion model. In practice, tackling class imbalance issue appropriately enables well-calibrated models to generate high-quality pseudo labels (see Figure 1 (c) & (d)), and in turn, improved quality of pseudo labels favorably manifests the mitigation of Matthew effect (see Figure 1 (a) & (b)), moving the learning toward unbiased.

Extensive experiments on various domains (*natural image domain*, *remote sensing image domain*, and *medical image domain*) across diverse benchmarks spanning different backbones demonstrate that our method performs favorably against state-of-the-art semi-supervised semantic segmentation methods, especially in the most limited label scenarios with the most severe class imbalance issues (*e.g.*, +2.6%/+2.0% compared to DDFP (Wang et al., 2024) and RankMatch (Mai et al., 2024) respectively on PASCAL *classic* under 1/16 protocol with the ResNet-101), evidencing the merits of modeling underlying distribution in the challenging dense pixel-level classification task.

2 Related Work

Class-Imbalanced Semi-Supervised Segmentation. Real-world datasets usually yield a class-imbalanced distribution, especially in dense prediction tasks (*e.g.*, semantic segmentation), making the standard training of machine learning models harder to generalize. Existing methods to re-balance the training objective can be roughly categorized into two paradigms: (1) re-sampling based methods (Chawla et al., 2002; He & Garcia, 2009; Byrd & Lipton, 2019; Chang et al., 2021; Shi et al., 2023; Wei et al., 2022) to adjust prediction labels by over-sampling the minority class or under-sampling the majority class. (2) re-weighting based methods (Cao et al., 2019; Cui et al., 2019; Huang et al., 2019; Ren et al., 2018; Hu et al., 2019; Chen et al., 2023d) to influence the loss function conditioned on specific criteria (*e.g.*, imposing the weights by strictly inverse the class frequency). However, all these methods assume all labels are accessible to alleviate the class imbalance issue and thus inapplicable to the unlabelled data in semi-supervised semantic segmentation. Recently, several studies have attempted to transfer these techniques on top of pseudo labels such as re-sampling (Wei et al., 2021), re-weighting (Wang et al., 2022a; Sun et al., 2023b; Xu et al., 2021; He et al., 2021; Wang et al., 2022b; Peng et al., 2023) (*e.g.*, Adsh (Guo & Li, 2022) utilizes adaptive thresholding that can be considered as binary weighting for semi-supervised learning, U^2PL (Wang et al., 2022b) adjusts the threshold adaptively to determine the reliability of pixels and constructs the extra supervised signal based on the negative classes of unreliable pixels, paying more attention to the tail classes), or a combination of both for semi-supervised learning (*e.g.*, AEL (Hu et al., 2021) adaptively balances the training of different categories). Nevertheless, these pseudo labels are often noisy as they are generated from poorly calibrated models. Furthermore, USRN (Guan et al., 2022) explores unbiased subclass regularization for alleviating the class imbalance issue. However, these discriminative methods are still confined to learning decision boundaries, which are brittle to the class imbalance issue, and the inherent nature of contempt for the underlying distribution remains unchanged. As a significant departure from the status quo, we formulate the semi-supervised semantic segmentation task as a conditional discrete data generation problem to model underlying distribution to overcome the shortcomings of discriminative solutions from a generative perspective.

Diffusion Models for Visual Perception. In addition to the significant progress in content generation, diffusion models have demonstrated incremental potential in the domain of perception (Chen et al., 2023b; Gu et al., 2022; Chen et al., 2023c; Brempong et al., 2022). Earlier studies primarily delve into investigating latent representations of diffusion models for zero-shot image segmentation (Baranchuk et al., 2021; Burgert et al., 2022) or applied diffusion models to medical image segmentation (Wolleb et al., 2022; Wu et al., 2022). Despite substantial progress, the outcomes of these efforts remain limited to specific local designs. The recent Pix2Seq-D (Chen et al., 2023c) extends the bit-diffusion (Chen et al., 2022) to panoptic segmentation, marking the first work of such expansion in a broader context. Additionally, DiffusionDet (Chen et al., 2023b) and DiffusionInst (Gu et al., 2022) explore diffusion models for query-based object detection (Carion et al., 2020) and instance segmentation (Zhang et al., 2021). Recently, several works have introduced diffusion into various semi-supervised tasks, such as classification, federated learning, time-series classification, and 3d object detection. Among them, both DPT (You et al., 2024) and FedDISC (Yang et al., 2024) aim to introduce an external diffusion model to generate data and utilize these data in a multi-stage training manner. DiffShape (Liu et al., 2024b) utilizes diffusion in a self-supervised manner to improve representation capability, and Diffusion-ss3d (Ho et al., 2023) exploits the denoising ability of the diffusion to improve the quality of the pseudo label. However, these methods differ from ours both from motivation to implementation. We comprehensively and meticulously compare our DiffMatch with these diffusion-based semi-supervised methods in Appendix F. In general, DiffMatch completely utilizes the characteristics of the diffusion process for semi-supervised semantic segmentation, aiming to provide a new perspective to alleviate the Matthew effect.

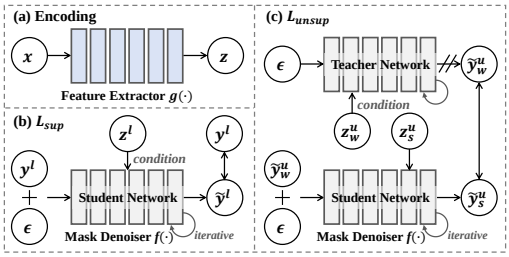


Figure 2: Our DiffMatch framework, which includes a feature extractor $g(\cdot)$ and a mask denoiser $f(\cdot)$. The diffusion process is conducted progressively in mask denoiser, aiming for lightweightness.

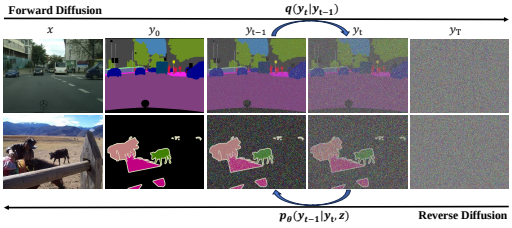


Figure 3: Conditional discrete data generation pipeline for semi-supervised semantic segmentation. Specifically, a conditional diffusion model is employed, where q is the forward diffusion process and p_θ is the inverse process.

3 METHOD

In this section, we first formulate the semi-supervised semantic segmentation problem as preliminaries (Section 3.1), specify the core framework of DiffMatch (Section 3.2 and Section 3.3), and elaborate the training (Section 3.4) and inference (Section 3.5) details.

3.1 PROBLEM DEFINITION

Given a labeled set $\mathcal{D}^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N^l}$ and an unlabeled set $\mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$, where N^l and N^u denote the number of labeled and unlabeled images, respectively, $N^u \gg N^l$, semi-supervised semantic segmentation aims to train a segmentation model with limited labeled data and fully exploit a large volume of unlabeled data. As shown in Figure 2, the popular teacher-student scheme consists of a teacher network and a student network. The student network is guided by two sources of supervision, including the ground truth $\boldsymbol{y}^l$ for the labeled data $\boldsymbol{x}_l$ (yielding supervised loss $\mathcal{L}_{sup}$) and the pseudo labels generated by the teacher network for the unlabeled data (constituting the unsupervised loss $\mathcal{L}_{unsup}$). In specific, for the unlabeled data, the unsupervised loss $\mathcal{L}_{unsup}$ is constructed in the form of consistency regularization, that is, the teacher network with a weakly augmented perturbation view $\boldsymbol{x}_w^u$ generates pseudo labels $\tilde{\boldsymbol{y}}_w^u$ to instruct the counterparts $\tilde{\boldsymbol{y}}_s^u$ from the student network under the presence of a strongly augmented perturbation view $\boldsymbol{x}_s^u$.

The teacher network can either be the same as the student network or an exponentially moving average (EMA) version of the student network. Note that in this paper, the teacher and student networks are identical, following UniMatch (Yang et al., 2022), to ensure simplicity and efficiency. The overall objective is the combination of supervised and unsupervised losses $\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{unsup}$.

In this work, we integrate generative models into semi-supervised learning, taking advantage of its better class-imbalance tolerance. In the next section, we elaborate on the modeling of our DiffMatch in detail, that is, how to realize closer collaboration between the diffusion process and the teacher-student paradigm.

3.2 THE DIFFMATCH FRAMEWORK

Figure 2 sheds light on the architecture of generation modeling for proposed DiffMatch. In specific, during training, the Gaussian noise $\boldsymbol{\epsilon}$ controlled by a noise schedule (Ho et al., 2020) is added to the ground truth $\boldsymbol{y}^l$ (from labeled data) or pseudo labels $\tilde{\boldsymbol{y}}_w^u$ (from unlabeled data) to construct the noisy masks. Then, the noisy mask is fused with the pixel embeddings $\boldsymbol{z}$ (acts as the condition) from the feature extractor $g(\cdot)$, and the resulting fused features are fed into a lightweight mask denoiser $f(\cdot)$ to generate the prediction without noise. At the inference phase, DiffMatch generates predictions by reversing the learned diffusion process, which transforms a known Gaussian distribution into a prediction that better matches the real distribution under the guidance of the images, maximizing the mutual information between the learned distribution and the underlying real one.

Due to the iterative nature of the diffusion sampling process, it requires multiple runs of the model during the inference phase. To minimize computational cost, we separate the entire network into two parts: the feature extractor and the mask denoiser following Chen et al. (2023c); Ji et al. (2023).

The former forward only once to extract the pixel embedding, and then the mask denoiser employs it as the condition rather than the raw image to iteratively reads out the prediction mask.

3.2.1 THE FEATURE EXTRACTOR

The feature extractor $g(\cdot)$ aims to extract the semantic features of the image $\boldsymbol{x}$ and upsample to a high-resolution pixel embedding $\boldsymbol{z} \in \mathbb{R}^{H \times W \times D}$ with an FPN-style structure (Lin et al., 2017) for sufficient representations. In our experiments, we adopt DeepLabv3+ (Chen et al., 2018) without the last layer classifier for a fair comparison (Na et al., 2023; Sun et al., 2023a), where $D = 256$.

3.2.2 THE MASK DENOISER

Denoiser Network. The input of the mask decoder $f(\cdot)$ is the concatenation of the noisy mask $\boldsymbol{y}_t$, which is obtained by adding Gaussian noise $\boldsymbol{\epsilon}$ to the ground truth from labeled data ($\boldsymbol{y}^l$) or pseudo labels from unlabeled data ($\tilde{\boldsymbol{y}}_w^u$), and the pixel embedding $\boldsymbol{z}$ from the feature extractor. To further minimize computational cost, we simply stack L layers of deformable attention (Zhu et al., 2020; Ji et al., 2023) as the mask denoiser (The number of layers L is set as 4 by default. Its effect can be referred to in Table 7). This lightweight design enables efficient reuse of shared parameters during multi-step denosing diffusion processes (*i.e.*, after running the feature extractor only once, reuse the efficient denoiser in several iterative steps), while maintaining highly competitive performance. More sophisticated mask denoiser are possible, to leverage recent advances in architecture designs (*e.g.*, TransUNet (Chen et al., 2021a)), but this is not our main focus so we opt for simplicity.

Forward and Reverse Process. Inspired by non-equilibrium thermodynamics, the optimization goal of the diffusion model is to maximize the likelihood to favor the alignment of the learned distribution and underlying real one. To this end, the diffusion model learns a series of state transitions (as shown in Figure 3) to transform noise $\boldsymbol{\epsilon}$ ($\boldsymbol{y}_T = \boldsymbol{\epsilon}$) from a known noise distribution into a data sample $\boldsymbol{y}_0$ from the data distribution $p(\boldsymbol{y}_0)$. To learn this mapping, we first define a forward transition $q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1})$ from state $\boldsymbol{y}_{t-1}$ to a more noisy state $\boldsymbol{y}_t$, which is defined as:

$$\boldsymbol{y}_t = \sqrt{\alpha_s}\boldsymbol{y}_{t-1} + \sqrt{1-\alpha_s}\boldsymbol{\epsilon} \implies q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}) = \mathcal{N}\left(\boldsymbol{y}_t; \sqrt{\alpha_s}\boldsymbol{y}_{t-1}, (1-\alpha_s)\mathbf{I}\right), \tag{1}$$

where t is from uniform density on $[0, 1]$ and $\boldsymbol{\epsilon}$ is drawn from standard normal density. α_s denotes the noise schedule (Ho et al., 2020; Song et al., 2020), meaning that the larger the time step t, the more the noise dominates and finally converges to pure Gaussian noise. Denoting the conditional reverse process as $p_\theta(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z})$, the straightforward objective is:

$$\mathcal{L}_{diff} = \sum_t D_{\mathrm{KL}}\left[q(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}) \| p_\theta(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z})\right]. \tag{2}$$

Benefiting from the reparameterization technique, the forward process can be simplified that directly obtaining $\boldsymbol{y}_t$ from $\boldsymbol{y}_0$, as:

$$\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \implies q(\boldsymbol{y}_t \mid \boldsymbol{y}_0) = \mathcal{N}\left(\boldsymbol{y}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{3}$$

where $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. Similarly, we can learn a mask denoiser $f(\cdot)$ to predict $\boldsymbol{y}_0$ directly from $\boldsymbol{y}_t$ under the guidance of $\boldsymbol{z}$, *i.e.*, $f(\boldsymbol{y}_t, \boldsymbol{z}) = p_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$. The objective can be simplified to:

$$\mathcal{L}_{diff} = \|f(\boldsymbol{y}_t, \boldsymbol{z}) - \boldsymbol{y}_0\|^2. \tag{4}$$

Note that in our semi-supervised setting, the data samples are either the ground truth mask from labeled data ($\boldsymbol{y}_0 = \boldsymbol{y}^l$) or pseudo labels from unlabeled data ($\boldsymbol{y}_0 = \tilde{\boldsymbol{y}}_w^u$). In specific, deriving from Equation 4, the supervised loss $\mathcal{L}_{sup}$ for labeled data can be formulated as:

$$\mathcal{L}_{sup} = \left\|f\left(\boldsymbol{y}_t^l, g(\boldsymbol{x}^l)\right) - \boldsymbol{y}_0^l\right\|^2. \tag{5}$$

In the same way, for the unlabeled data, the unsupervised loss $\mathcal{L}_{unsup}$ can be formulated as:

$$\mathcal{L}_{unsup} = \left\|f\left(\widetilde{\boldsymbol{y}}_{t,w}^u, g(\boldsymbol{x}_s^u)\right) - \widetilde{\boldsymbol{y}}_{0,w}^u\right\|^2, \tag{6}$$

where $\widetilde{\boldsymbol{y}}_{0,w}^u = f(\boldsymbol{\epsilon}, g(\boldsymbol{x}_w^u))$ denotes the pseudo labels and s/w means the strong/weak augmentation. Intuitively, the unsupervised loss fits with the consistency regularization of a standard teacher-student paradigm in semi-supervised semantic segmentation. In Algorithm 1, we present the pseudo algorithm of DiffMatch to clearly summarize our method. At this point, we have explored the integration of the diffusion process and the teacher-student paradigm to alleviate the class imbalance issue from a generative perspective.

3.3 DEBIASED ADJUSTMENT

Given the long-tailed nature of the class distribution $p(\boldsymbol{y}_0)$ in practice, the learned conditional inverse probability $p_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$ is inevitably biased. To further improve the tolerance of the diffusion model to class imbalance, we propose the debiased adjustment. First, we represent the conditional inverse probability under ideal condition $*$ (*i.e.*, when the class distribution is uniform, $p^*(\boldsymbol{y}_0) = \frac{1}{C}$, where C is the number of classes) as $p^*_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$. With the Bayesian formula, we deduce the relation between $p_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$ and $p^*_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$ (refer to Appendix C for detailed derivation):

$$p^*_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) = p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) \frac{p_\theta\left(\boldsymbol{y}_t\right)}{p^*_\theta\left(\boldsymbol{y}_t\right)} \frac{\hat{q}^*\left(\boldsymbol{y}_{t+1}\right)}{\hat{q}\left(\boldsymbol{y}_{t+1}\right)}. \tag{7}$$

Intuitively, we can obtain the ideal conditional inverse probability $p^*_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)$ by modulate $p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)$ by a factor $\frac{p_\theta(\boldsymbol{y}_t)}{p^*_\theta(\boldsymbol{y}_t)} \frac{\hat{q}^*(\boldsymbol{y}_{t+1})}{\hat{q}(\boldsymbol{y}_{t+1})}$. However, directly estimating this modulation factor at each time step t is highly challenging. Instead, we incorporate it into the training loss function to achieve an equivalent objective. Replacing the $p_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$ in Equation 2 with $p^*_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$:

$$\begin{aligned}
\mathcal{L}^*_{diff} =& \sum_t D_{\mathrm{KL}}\left[q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}\right) \| p^*_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)\right] \\
=& \sum_t \left\{ D_{\mathrm{KL}}\left[q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}\right) \| p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)\right] + t D_{\mathrm{KL}}\left[\frac{p_\theta\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t\right)}{C p_\theta\left(\boldsymbol{y}_0\right)} \| p^*_\theta\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t\right)\right] \right\} \\
=& \mathcal{L}_{diff} + \sum_t \left\{ t D_{\mathrm{KL}}\left[\frac{p_\theta\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t\right)}{C p_\theta\left(\boldsymbol{y}_0\right)} \| p^*_\theta\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t\right)\right] \right\}.
\end{aligned} \tag{8}$$

In practice, we approximate the $p_\theta(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t)$ with Monte-Carlo sampling from $p_\theta(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{z})$ and the loss reduces to:

$$\mathcal{L}^*_{diff} = \left\| f\left(\boldsymbol{y}_t, \boldsymbol{z}\right) - \boldsymbol{y}_0 \right\|^2 + \tau t \left\| f\left(\boldsymbol{y}_t, \boldsymbol{z}\right) - \frac{f\left(\boldsymbol{y}_t, \boldsymbol{z}\right)}{C p\left(\boldsymbol{y}_0\right)} \right\|^2, \tag{9}$$

where τ is the trade-off weight for the regularization term, set to 0.1 by default, and C is the number of classes. Please refer to Appendix C for detailed derivation. Intuitively, the second term imposes a constraint directly between the prediction of mask denoiser and its roughly debiased version, reducing the risk of overfitting to the head classes and increasing coverage of the tail class distribution. Based on Equation 9, the supervised loss and unsupervised loss are updated as:

$$\mathcal{L}_{sup} = \left\| f\left(\boldsymbol{y}^l_t, g(\boldsymbol{x}^l)\right) - \boldsymbol{y}^l_0 \right\|^2 + \tau t \left\| f\left(\boldsymbol{y}^l_t, g(\boldsymbol{x}^l)\right) - \frac{f\left(\boldsymbol{y}^l_t, g(\boldsymbol{x}^l)\right)}{C p\left(\boldsymbol{y}^l_0\right)} \right\|^2, \tag{10}$$

$$\mathcal{L}_{unsup} = \left\| f\left(\widetilde{\boldsymbol{y}}^u_{t,w}, g(\boldsymbol{x}^u_s)\right) - \widetilde{\boldsymbol{y}}^u_{0,w} \right\|^2 + \tau t \left\| f\left(\widetilde{\boldsymbol{y}}^u_{t,w}, g(\boldsymbol{x}^u_s)\right) - \frac{f\left(\widetilde{\boldsymbol{y}}^u_{t,w}, g(\boldsymbol{x}^u_s)\right)}{C p\left(\widetilde{\boldsymbol{y}}^u_{0,w}\right)} \right\|^2. \tag{11}$$

Note that, in our implementation, $p(\boldsymbol{y}^l_0)$ is statistically derived from the ground truth of labeled data while the $p(\widetilde{\boldsymbol{y}}^u_{0,w})$ is initialized as $p(\boldsymbol{y}^l_0)$ and updated based on its own pseudo label in an exponential moving average (EMA) manner to progressively align the class prior on unlabeled data. By formulating the pseudo label generation of consistency regularization as an optimization problem progressively solved by the denoising diffusion process, DiffMatch bridges the gap by drifting biased prediction towards unbiased learning.

3.4 TRAINING

Our main training objective is to learn a series of state transitions under the guidance of input image to transform noise from a known noise distribution into prediction that better matches real class distribution. We adopt analog bits encoding strategy (Chen et al., 2022) to first convert discrete integers from ground truth or pseudo labels into bit strings, and then cast them as real number. When constructing the analog bits, we can shift and scale them into $\{-b, b\}$ (The scaling factor b is by default set to 0.1. Its impact can be referred to in Table 8). To draw samples, we follow the same procedure as sampling in a continuous diffusion model, except that we apply a quantization operation at the end by simply thresholding the generated analog bits. The training procedure for the diffusion process is provided in Algorithm 2.

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

Method	ResNet-50					ResNet-101				
	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
Sup.-only	44.0	52.3	61.7	66.7	72.9	45.1	55.3	64.8	69.7	73.5
FixMatch[NeurIPS'20]	60.1	67.3	71.4	73.7	76.9	63.9	73.0	75.5	77.8	79.2
PCR[NeurIPS'22]	–	–	–	–	–	70.0	74.7	77.1	78.5	80.7
GTA-Seg[NeurIPS'22]	–	–	–	–	–	70.0	73.2	75.6	78.4	80.5
ReCo[ICLR'22]	64.8	72.0	73.1	74.7	–	–	–	–	–	–
AugSeg[CVPR'23]	64.2	72.1	76.1	77.4	78.8	71.0	75.4	78.8	80.3	81.3
UniMatch[CVPR'23]	67.4	71.9	75.3	78.0	79.3	73.5	75.4	78.7	80.2	81.9
NP-SemiSeg[ICML'23]	65.7	72.3	75.7	77.4	–	–	–	–	–	–
DAW[NeurIPS'23]	68.5	73.1	76.3	78.6	79.7	74.8	77.4	79.5	80.6	81.5
DDFP[CVPR'24]	–	–	–	–	–	74.9	78.0	79.5	81.2	81.9
RankMatch[CVPR'24]	71.6	74.6	76.7	78.8	80.0	75.5	77.6	79.8	80.7	82.2
PRCL[IJCV'24]	–	–	–	–	–	71.2	72.2	75.2	76.2	78.3
DiffMatch (Ours)	**73.3**	**75.7**	**77.9**	**79.6**	**81.6**	**77.5**	**78.3**	**80.6**	**81.5**	**83.3**
$\Delta\uparrow$	+29.3	+23.4	+16.2	+12.9	+8.7	+32.4	+23.0	+15.8	+11.8	+9.8

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

Method	ResNet-50			ResNet-101		
	1/16 (662)	1/8 (1323)	1/4 (2646)	1/16 (662)	1/8 (1323)	1/4 (2646)
Sup.-only	62.4	68.2	72.3	67.5	71.1	74.2
FixMatch[NeurIPS'20]	70.6	73.9	75.1	74.3	76.3	76.9
AEL[NeurIPS'21]	–	–	–	77.2	77.6	78.1
PCR[NeurIPS'22]	–	–	–	78.6	80.7	80.8
GTA-Seg[NeurIPS'22]	–	–	–	77.8	80.5	80.6
AugSeg[CVPR'23]	74.7	76.0	77.2	77.0	77.3	78.8
UniMatch[CVPR'23]	75.8	76.9	76.8	78.1	78.4	79.2
CFCG[IJCCV'23]	75.0	77.1	77.7	76.8	79.1	79.9
NP-SemiSeg[ICML'23]	73.4	76.5	76.7	–	–	–
DAW[NeurIPS'23]	76.2	77.6	77.4	78.5	78.9	79.6
DDFP[CVPR'24]	–	–	–	78.3	78.8	79.8
RankMatch[CVPR'24]	76.6	77.8	78.3	78.9	79.2	80.0
PRCL[IJCV'24]	–	–	–	77.9	79.1	79.9
DiffMatch (Ours)	**77.9**	**78.7**	**79.0**	**80.3**	**81.4**	**81.6**
$\Delta\uparrow$	+15.5	+10.5	+6.7	+12.8	+10.3	+7.4

3.5 Inference

At the inference phase, the target data sample $\boldsymbol{y}_0$ is reconstructed from noise $\boldsymbol{y}_T$ with the mask denoiser $f(\cdot)$ and an updating rule (Song et al., 2020; Ho et al., 2020) in an iterative Markovian way. We choose the DDIM update rule (Song et al., 2020) for the sampling process. We also represent the trade-off between performance and computation by different sampling steps for multi-step inference in Table 5. Please refer to Algorithm 3 for details about the sampling procedure for diffusion process. Note that to reduce inference overhead, we do not employ any post-processing techniques, such as self-condition (Chen et al., 2022), and sampling drift (Ji et al., 2023), *etc*.

4 Experiments

In this section, we give comprehensive evaluations of various class-imbalanced datasets. We first describe the experimental setups in Section 4.1. Then, we present the empirical results of our DiffMatch and other compared competitors under extensive setups in Section 4.2. Finally, we present detailed analyses to help understand our method in Section 4.3.

4.1 Experimental Setup

Datasets. We conduct experiments on three datasets with severe class-imbalanced issues. (1) **PASCAL VOC 2012** (Everingham et al., 2010) contains 21 classes with 1,464 and 1,449 finely annotated images for training and validation, respectively. We augment the original training set (*i.e.*, *classic*) with additional 9,118 coarsely annotated images in SBD (Hariharan et al., 2011) to get a *blender* training set following other researches (Chen et al., 2021b; Hu et al., 2021). According to statistics,

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

Method	ResNet-50				ResNet-101			
	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
Sup.-only	63.3	70.2	73.1	76.6	66.3	72.8	75.0	78.0
FixMatch[NeurIPS'20]	72.6	75.7	76.8	78.2	74.2	76.2	77.2	78.4
AEL[NeurIPS'21]	74.0	75.8	76.2	–	75.8	77.9	79.0	80.3
PCR[NeurIPS'22]	–	–	–	–	73.4	76.3	78.4	79.1
GTA-Seg[NeurIPS'22]	–	–	–	–	69.4	72.0	76.1	–
AugSeg[CVPR'23]	73.7	76.5	78.8	79.3	75.2	77.8	79.5	80.4
UniMatch[CVPR'23]	75.0	76.8	77.5	78.6	76.6	77.9	79.2	79.5
Co-Train[ICCV'23]	–	76.3	77.1	–	75.0	77.3	78.7	–
NP-SemiSeg[ICML'23]	73.0	77.1	78.8	78.7	–	–	–	–
DAW[NeurIPS'23]	75.2	77.5	79.1	79.5	76.6	78.4	79.8	80.6
DDFP[CVPR'24]	–	–	–	–	77.1	78.1	79.8	80.8
RankMatch[CVPR'24]	75.4	77.7	79.2	79.5	77.1	78.6	80.0	80.7
PRCL[IJCV'24]	–	–	–	–	73.4	77.0	77.9	80.0
DiffMatch (Ours)	**76.5**	**78.3**	**79.8**	**80.0**	**77.8**	**79.1**	**80.5**	**81.3**
$\Delta\uparrow$	**+13.2**	**+8.1**	**+6.7**	**+3.4**	**+11.5**	**+6.3**	**+5.5**	**+3.3**

the pixel number of the head class *background* is more than 200× that of the tail class *bicycle*. (2) **Cityscapes** (Cordts et al., 2016) consists of 2,975 images for training and 500 images for validation with 19 classes. The ratio of head class *road* to tail class *motorcycle* reaches 400. (3) **COCO** (Lin et al., 2014), composed of 118k/5k training/validation images, is a more severe class-imbalanced dataset, containing 81 classes to predict, with over 10,000 head-to-tail ratio. To further demonstrate the versatility of DiffMatch, we extend our experiments in two crucial real-world applications: remote sensing interpretation and medical image analysis, as shown in Appendix J

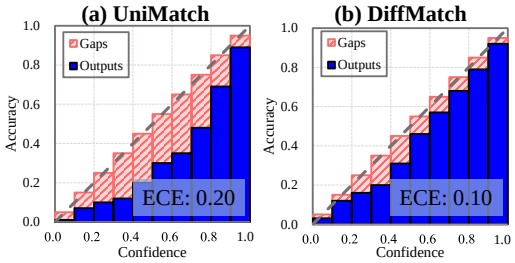


Figure 4: Calibration on unlabeled data produced by UniMatch (left) and DiffMatch (right).

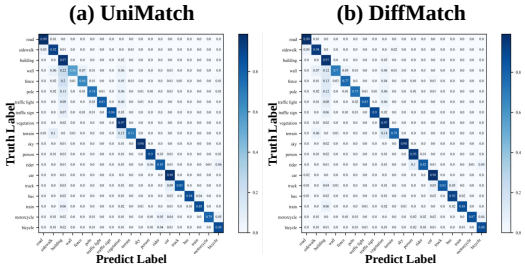


Figure 5: Confusion matrix on unlabeled data of UniMatch (left) and DiffMatch (right).

Implementation Details. For a fair and exhaustive comparison, we use ResNet-50/101 (He et al., 2016b) pretrained on ImageNet (Krizhevsky et al., 2012) and Xception-65 (Chollet, 2017) as the backbones and DeepLabv3+ (Chen et al., 2018) as the decoder. We set the sampling step as 3 at inference, the number of layers in mask denoiser L as 4 and the scaling factor b as 0.1 for all experiments. During training, we randomly crop 513×513 for PASCAL and COCO datasets, and train 80 and 30 epochs, respectively. For Cityscapes, the cropsize is set as 801×801 and the training epoch is 240. The batch size of the three datasets is set to 8. Polynomial Decay learning rate policy is applied throughout the whole training. The strong augmentation contains feature dropout, random color jitter, grayscale and Gaussian blur. The weak augmentation consists of random crop, resize and horizontal flip. All experiments are conducted on 8× RTX 3090 GPUs (memory is 24G/GPU).

4.2 Empirical Results

We evaluate our method on PASCAL (*classic* and *blender*), Cityscapes datasets with ResNet-50/101, and COCO dataset with Xception-65 under different semi-supervised learning settings (*i.e.*, partition protocols). The partition protocol (*e.g.*, 1/16) indicates the ratio of labeled data used in training to the entire training set. It is worth noting that the smaller the partition protocol, the less labeled data is used for training, and the more biased the training may be. The consistently dominant performance under all partition protocols with different backbones on all datasets against other competitors (FixMatch (Sohn et al., 2020), PseudoSeg (Zou et al., 2020), AEL (Hu et al., 2021), ReCo (Liu et al., 2021), PC2Seg (Zhong et al., 2021), PCR (Xu et al., 2022), GTA-Seg (Jin et al., 2022),

Table 4: Quantitative results of different SSL methods on COCO.

Method	1/512	1/256	1/128	1/64	1/32
Sup.-only	22.9	28.0	33.6	37.8	42.2
PseudoSeg	29.8	37.1	39.1	41.8	43.6
PC2Seg	29.9	37.5	40.1	43.7	46.1
MKD	30.2	38.0	42.3	45.5	47.3
UniMatch	31.9	38.9	44.4	48.2	49.8
DiffMatch (Ours)	**34.6**	**41.9**	**47.2**	**49.8**	**52.4**
Δ↑	+11.7	+13.9	+13.6	+12.0	+10.2

Table 5: Accuracy *vs.* Efficiency.

Method		mIoU (92)	mIoU (1464)	FPS (↑)	#Param
UniMatch		67.4	79.3	24.9	40.5M
Dis. Baseline		67.9	79.5	23.8	44.9M
DiffMatch w/o *adj.*		72.2	81.3	19.8	44.9M
DiffMatch	step1	68.7	79.9	23.3	44.9M
	step2	71.2	80.7	21.2	
	step3	**73.3**	**81.6**	19.6	
	step4	73.3	81.4	18.2	
	step5	73.4	81.7	16.9	

Table 6: Performance of head & tail classes.

ResNet-50	PASCAL *classic* 1/16 (92) mIoU	mIoU$_h$	mIoU$_t$
Sup.-only	44.0	66.5	28.1
♦ Re-Sampling	45.6	67.8	29.3
♦ Re-weighting	46.2	68.3	30.1
♠ FixMatch	60.1	78.4	48.4
♠ ReCo	64.8	81.2	49.6
♠ NP-SemiSeg	65.8	82.7	50.2
♣ DARP	61.5	79.9	49.0
♣ CReST	62.2	80.6	49.4
♣ FreeMatch	62.3	80.2	49.1
♣ DARS	62.7	82.5	50.3
♣ AEL	66.3	84.2	51.1
♣ U^{2}PL	67.4	85.3	53.7
♣ USRN	66.8	83.9	51.8
DiffMatch (Ours)	**73.3**	**89.3**	**66.8**

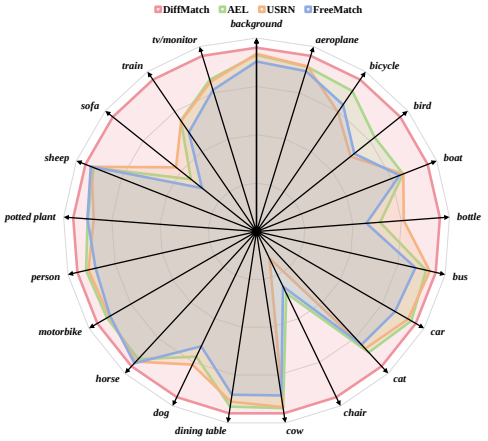


Figure 6: Normalized performance on PASCAL *classic* 92 for per class.

UniMatch (Yang et al., 2022), AugSeg (Zhao et al., 2023c), NP-SemiSeg (Wang et al., 2023c), DAW (Sun et al., 2023b), CFCG (Li et al., 2023a), Co-Train (Li et al., 2023b), MKD (Yuan et al., 2023), DDFP (Wang et al., 2024), RankMatch (Mai et al., 2024), PRCL (Xie et al., 2024)) proves the effectiveness of our DiffMatch against the class imbalance issue, evidencing the merits of modeling underlying distribution in the challenging dense pixel-level classification task.

Results on PASCAL. Table 1 and Table 2 show the comparison of our method with the SOTA methods on PASCAL *classic* and *blender* set. Compared with the supervised-only (*Sup.-only*) model, our method achieves considerable performance improvements, demonstrating that the information in unlabeled data is effectively utilized in our method. Moreover, in the label-scarce scenario, *e.g.*, 1/16 (92) in PASCAL *classic*, our approach achieves 73.3% and 77.5% mIoU with the backbone ResNet-50 and ResNet-101, boosting the SOTA DAW (Sun et al., 2023a) by 4.8% and 2.7%, respectively. These superior results prove that our training is more unbiased.

Results on Cityscapes. Table 3 summarizes the performance of our DiffMatch and compared methods on the Cityscapes dataset. For the more class-imbalanced dataset (the ratio of head class road to tail class motorcycle reaches 400), our method still achieves SOTA performance. Specifically, compared with the leading methods DAW (Sun et al., 2023a), DiffMatch improves up to 1.3%/1.2% at absolute mIoU gain under 1/16 partition protocols with ResNet-50/ResNet-101, respectively, showing the superiority of our method over discriminative methods.

Results on COCO. COCO is a large-scale dataset where the class imbalance issue is most severe (the number of head-to-tail ratio is more than 10,000). In Table 4, DiffMatch achieves surprising performance lift compared with the discriminative model. For example, under the 1/512 partition protocol, the performance of DiffMatch is superior to that of UniMatch (Yang et al., 2022) (34.6% *vs.* 31.9%), this is in line with the goal of DiffMatch against class imbalance issue.

4.3 Detailed Analyses

Performance in Head&Tail Classes. Considering that the Matthew effect refers to the bias in model predictions, it can also be viewed as a measure of model calibration. This directly impacts

the quality of pseudo-labels for unlabeled data, thereby affecting the model's performance across different classes. Therefore, we compare DiffMatch to other competitive methods to analyze the effectiveness of addressing class imbalance by examining the performance of head/tail classes. To show the source of our absolute performance gain, we present the mIoU of the top-5 classes (mIoU_h) and the bottom-5 classes (mIoU_t) under PASCAL *classic* 1/16 (92) with ResNet-50. To make a comprehensive comparison, we reproduce several methods based on their open-source code under the same experiment setting, including class-imbalanced learning methods ♦, SSL methods ♠, and recently proposed class-imbalanced SSL methods ♣. (1) Specifically, for class-imbalanced learning, we consider the two most popular paradigms: a) Re-Sampling (Byrd & Lipton, 2019) and b) Re-weighting (Cui et al., 2019); (2) for SSL methods, we take Fixmatch (Sohn et al., 2020), ReCo (Liu et al., 2021) and NP-SemiSeg (Wang et al., 2023c) into consideration. (3) To further show the efficacy of our proposal, we also compare it with the recently proposed algorithms that consider SSL and class imbalance issues simultaneously, including DARP (Kim et al., 2020), CReST (Wei et al., 2021), FreeMatch (Wang et al., 2022a), DARS (He et al., 2021), AEL (Hu et al., 2021), U^2PL (Wang et al., 2022b) and USRN (Guan et al., 2022). Please refer to Section 2 for more details.

As depicted in Table 6, we have the following findings: (1) It is not desirable to directly apply the class-imbalanced learning method to SSL tasks because it does not utilize unlabeled data. (2) SSL methods achieve certain performance gains, but still underperform in the tail classes. (3) Thanks to the modeling of distributions and the derived debiased adjustment , DiffMatch yields favorable performance especially in the tail classes, effectively alleviating the Matthew effect. To better understand the prediction bias of each class, as Figure 6 illustrates, DiffMatch achieves more unbiased predictions on all 21 classes. Moreover, we provide training curves for the number of pseudo labels in the head (*road*) and tail (*motorcycle*) classes in the Appendix D, demonstrating the effectiveness of our DiffMatch in mitigating the Matthew effect.

Accuracy *vs.* Efficiency. We show the dynamic trade-off of DiffMatch between accuracy and efficiency in Table 5. To begin with, we construct a discriminative baseline (*Dis.* Baseline) with the same extra deformable attention layers. (1) Comparing the 1st and 2nd rows, we can see that simply increasing the number of parameters in the model does not lead to an effective performance improvement. Then, 2nd *vs.* 3rd indicates that the performance improvement of DiffMatch primarily stems from modeling the underlying distribution, as opposed to discriminative models (*Dis.* Baseline). (2) Comparing the 3nd row (DiffMatch w/o *adj.*) and the final DiffMatch, we can observe a clear performance lift credited to debiased adjustment. This suggests the effectiveness of debiased adjustment to adjust the conditional reverse probability, reducing the risk of overfitting to the head classes and increasing coverage of the tail class distribution. (3) With the sampling step increase, the performance gets better (same result can also be observed in Figure 10). When adopting 3 sampling steps, the performance is further boosted while maintaining comparable FPS. These results show that DiffMatch can iteratively infer multiple times with reasonable time cost.

Quality of Pseudo Label. To take a close look at DiffMatch, we showcase the confusion matrix (Figure 5) and expected calibration error (Figure 4) on unlabeled data to directly measure the performance of different models in the Matthew effect and model calibration respectively, on the $1/16$ partition protocol of the Cityscapes dataset. The results show that the raw pseudo-labels generated by UniMatch are biased toward the majority classes. For example, there are more than 20% examples that belong to class *wall* are predicted wrongly as class *building*. On the contrary, our DiffMatch can achieve a more unbiased confusion matrix, striving to mitigate the Matthew effect. These results indicate that the quality of pseudo-labels is actually improved, which can help to improve the generalization performance. Similarly, a better-calibrated model is obtained thanks to the modeling of the underlying distribution by our DiffMatch (Figure 4). Based on this, well-calibrated models will generate high-quality pseudo labels, and in turn, improved quality of pseudo labels could result in a better estimation of distribution.

5 Conclusion

In this paper, we analyze the Matthew effect in previous methods that hinder model learning when dealing with class imbalance issues from a discriminative view. we propose DiffMatch to formulate the semi-supervised semantic segmentation task as a conditional discrete data generation problem to model underlying distribution against the Matthew effect. DiffMatch offers a fresh generative perspective to alleviating class imbalance, and we believe it has the potential to complement other semi-supervised learning strategies to facilitate future advancements.

REFERENCES

Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2021.

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11514–11524, 2023.

Wele Gedara Chaminda Bandara and Vishal M Patel. Revisiting consistency regularization for semi-supervised change detection in remote sensing images. *arXiv preprint arXiv:2204.08454*, 2022.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Emmanuel Asiedu Brempong, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Denoising pretraining for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4175–4186, 2022.

Ryan Burgert, Kanchana Ranasinghe, Xiang Li, and Michael S Ryoo. Peekaboo: Text to image diffusion models are zero-shot segmentors. *arXiv preprint arXiv:2211.13224*, 2022.

Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pp. 872–881. PMLR, 2019.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Animashree Anandkumar, Sanja Fidler, and Jose M Alvarez. Image-level or object-level? a tale of two resampling strategies for long-tailed detection. In *International conference on machine learning*, pp. 1463–1472. PMLR, 2021.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023a.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021a.

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)*, pp. 801–818, 2018.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19830–19843, 2023b.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 909–919, 2023c.

Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19277–19287, 2023d.

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*, pp. 2613–2622, 2021b.

Zenggui Chen and Zhouhui Lian. Semi-supervised semantic segmentation via prototypical contrastive learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 6696–6705, 2022.

Hanyang Chi, Jian Pang, Bingfeng Zhang, and Weifeng Liu. Adaptive bidirectional displacement for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4070–4080, 2024.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

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*, pp. 3213–3223, 2016.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Zhangxuan Gu, Haoxing Chen, Zhuoer Xu, Jun Lan, Changhua Meng, and Weiqiang Wang. Diffusioninst: Diffusion model for instance segmentation. *arXiv preprint arXiv:2212.02773*, 2022.

Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Unbiased subclass regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9968–9978, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Lan-Zhe Guo and Yu-Feng Li. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International Conference on Machine Learning*, pp. 8082–8094. PMLR, 2022.

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

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016a.

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*, pp. 770–778, 2016b.

Ruifei He, Jihan Yang, and Xiaojuan Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6930–6940, 2021.

Cheng-Ju Ho, Chen-Hsuan Tai, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Diffusion-ss3d: Diffusion model for semi-supervised 3d object detection. *Advances in Neural Information Processing Systems*, 36:49100–49112, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Prantik Howlader, Srijan Das, Hieu Le, and Dimitris Samaras. Beyond pixels: Semi-supervised semantic segmentation with a multi-scale patch-based multi-label classifier. In *European Conference on Computer Vision*, pp. 342–360. Springer, 2025a.

Prantik Howlader, Hieu Le, and Dimitris Samaras. Weighting pseudo-labels via high-activation feature index similarity and object detection for semi-supervised segmentation. In *European Conference on Computer Vision*, pp. 456–474. Springer, 2025b.

Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. Semivl: semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pp. 257–275. Springer, 2025.

Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021.

Zhiting Hu, Bowen Tan, Russ R Salakhutdinov, Tom M Mitchell, and Eric P Xing. Learning data manipulation for augmentation and weighting. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2781–2794, 2019.

J Igelsias, M Styner, T Langerak, B Landman, Z Xu, and A Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21741–21752, 2023.

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.

Jaehyung Kim, Youngbum Hur, Sejun Park, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning. *Advances in neural information processing systems*, 33:14567–14579, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1205–1214, 2021.

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

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*, volume 3, pp. 896, 2013.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323, 2021.

Shuo Li, Yue He, Weiming Zhang, Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, and Jingdong Wang. Cfcg: Semi-supervised semantic segmentation via cross-fusion and contour guidance supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16348–16358, 2023a.

Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*, 2023b.

Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16197–16208, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Qianying Liu, Xiao Gu, Paul Henderson, and Fani Deligianni. Multi-scale cross contrastive learning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2306.14293*, 2023.

Qianying Liu, Paul Henderson, Xiao Gu, Hang Dai, and Fani Deligianni. Learning semi-supervised medical image segmentation from spatial registration. *arXiv preprint arXiv:2409.10422*, 2024a.

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

Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18(5):811–815, 2020.

Zhen Liu, Wenbin Pei, Disen Lan, and Qianli Ma. Diffusion language-shapelets for semi-supervised time-series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14079–14087, 2024b.

Xiangde Luo, Minhao Hu, Tao Song, Guotai Wang, and Shaoting Zhang. Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In *International conference on medical imaging with deep learning*, pp. 820–833. PMLR, 2022.

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*, pp. 3391–3401, 2024.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379, 2019.

Jaemin Na, Jung-Woo Ha, Hyung Jin Chang, Dongyoon Han, and Wonjun Hwang. Switching temporary teachers for semi-supervised semantic segmentation. *arXiv preprint arXiv:2310.18640*, 2023.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

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

Daifeng Peng, Lorenzo Bruzzone, Yongjun Zhang, Haiyan Guan, Haiyong Ding, and Xu Huang. Semicdnet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7): 5891–5906, 2020.

Hanyu Peng, Weiguo Pian, Mingming Sun, and Ping Li. Dynamic re-weighting for long-tailed semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6464–6474, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343. PMLR, 2018.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

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.

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? *Advances in Neural Information Processing Systems*, 36, 2023.

Mennatullah Siam, Mostafa Gamal, Moemen Abdel-Razek, Senthil Yogamani, Martin Jagersand, and Hong Zhang. A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 587–597, 2018.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3097–3107, 2024.

Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: Exploring the better weighting function for semi-supervised semantic segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Rui Sun, Huayu Mai, Tianzhu Zhang, and Feng Wu. Daw: Exploring the better weighting function for semi-supervised semantic segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740, 2022.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2021.

Changqi Wang, Haoyu Xie, Yuhui Yuan, Chong Fu, and Xiangyu Yue. Space engage: Collaborative space supervision for contrastive-based semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 931–942, 2023a.

Jia-Xin Wang, Si-Bao Chen, Chris HQ Ding, Jin Tang, and Bin Luo. Ranpaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021.

Jianfeng Wang, Daniela Massiceti, Xiaolin Hu, Vladimir Pavlovic, and Thomas Lukasiewicz. Np-semiseg: when neural processes meet semi-supervised semantic segmentation. In *International Conference on Machine Learning*, pp. 36138–36156. PMLR, 2023b.

Jianfeng Wang, Daniela Massiceti, Xiaolin Hu, Vladimir Pavlovic, and Thomas Lukasiewicz. Np-semiseg: when neural processes meet semi-supervised semantic segmentation. In *International Conference on Machine Learning*, pp. 36138–36156. PMLR, 2023c.

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*, pp. 3303–3312, 2024.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Zhen Wu, and Jindong Wang. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022a.

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*, pp. 4248–4257, 2022b.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10857–10866, 2021.

Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *International Conference on Machine Learning*, pp. 23615–23630. PMLR, 2022.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pp. 1336–1348. PMLR, 2022.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. *arXiv preprint arXiv:2211.00611*, 2022.

Linshan Wu, Leyuan Fang, Xingxin He, Min He, Jiayi Ma, and Zhun Zhong. Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Haoyu Xie, Changqi Wang, Jian Zhao, Yang Liu, Jun Dan, Chong Fu, and Baigui Sun. Prcl: Probabilistic representation contrastive learning for semi-supervised semantic segmentation. *International Journal of Computer Vision*, pp. 1–19, 2024.

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.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pp. 11525–11536. PMLR, 2021.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. *arXiv preprint arXiv:2208.09910*, 2022.

Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. Exploring one-shot semi-supervised federated learning with pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16325–16333, 2024.

Zebin You, Yong Zhong, Fan Bao, Jiacheng Sun, Chongxuan Li, and Jun Zhu. Diffusion models and semi-supervised learners benefit mutually with few labels. *Advances in Neural Information Processing Systems*, 36, 2024.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22*, pp. 605–613. Springer, 2019.

Jianlong Yuan, Jinchao Ge, Zhibin Wang, and Yifan Liu. Semi-supervised semantic segmentation with mutual knowledge distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5436–5444, 2023.

Shiying Yuan, Ruofei Zhong, Cankun Yang, Qingyang Li, and YaXin Dong. Dynamically updated semi-supervised change detection network combining cross-supervision and screening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5729–5739, 2023a.

Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. *arXiv preprint arXiv:2211.11335*, 2022a.

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. *arXiv preprint arXiv:2212.04976*, 2022b.

Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23705–23714, 2023b.

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*, pp. 11350–11359, 2023c.

Zhen Zhao, Zicheng Wang, Longyue Wang, Dian Yu, Yixuan Yuan, and Luping Zhou. Alternate diverse teaching for semi-supervised medical image segmentation. In *European Conference on Computer Vision*, pp. 227–243. Springer, 2025.

Shuhong Zheng, Zhipeng Bao, Ruoyu Zhao, Martial Hebert, and Yu-Xiong Wang. Diff-2-in-1: Bridging generation and dense perception with diffusion models. *arXiv preprint arXiv:2411.05005*, 2024.

Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7273–7282, 2021.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.

A MOTIVATION FOR GENERATIVE MODELS TO ALLEVIATE CLASS IMBALANCE ISSUE

Previous research (Ng & Jordan, 2001) theoretically derived the differences in generalization error $\varepsilon(\cdot)$ between discriminative (Dis) and generative models (Gen) under ideal conditions (*i.e.*, with an infinite number of samples ∞), where m denotes the number of samples, n is the number of model parameters, and $G(\cdot)$ represents a small meaningful bound.

$$\varepsilon\left(h_{\text{Dis}}\right) \leq \varepsilon\left(h_{\text{Dis},\infty}\right) + O\left(\sqrt{\frac{n}{m}\log\frac{m}{n}}\right) \tag{12}$$

$$\varepsilon\left(h_{\text{Gen}}\right) \leq \varepsilon\left(h_{\text{Gen},\infty}\right) + G\left(O\left(\sqrt{\frac{1}{m}\log n}\right)\right) \tag{13}$$

The above theory demonstrates that the *asymptotic error approaching rate* of generative models is $O(\log n)$, which is better than the discriminative model's ($O(n)$). In other words, under the same number of model parameters, generative models can approach the optimal form under the ideal condition (*i.e.*, infinite training sample) with fewer training samples (logarithmic number, *i.e.*, $O(\log n)$), compared to the discriminative model, which requires a linear number of samples ($O(n)$). This provides a special bonus for the inherent class imbalance problem in semi-supervised semantic segmentation, particularly for tail classes. Specifically, generative models have better potential to enable tail classes with extremely limited sample quantity to converge to the form assumed under sufficient sample conditions, conceptually bridging the gap with the ample samples of head classes, *i.e.*, *better class-imbalance tolerance*.

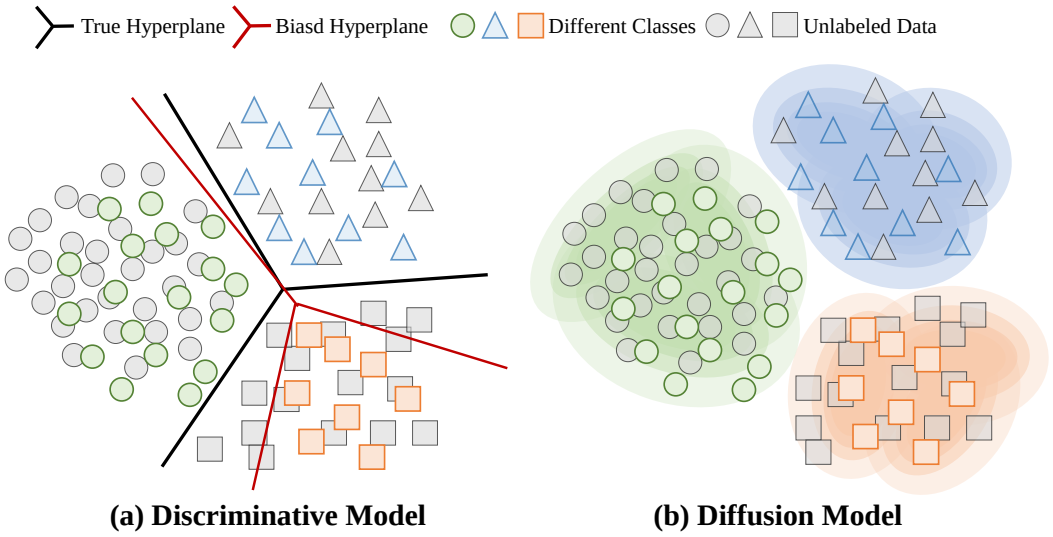


Figure 7: Illustration on discriminative model *vs.* diffusion model.

On the other hand, from the perspective of optimization objectives, diffusion-based generative models and discriminative models have fundamentally different optimization objectives. Specifically, discriminative models are typically trained by minimizing empirical risk, aiming to minimize the prediction error or loss function of the model solely on the training data. In this case, these methods, only learning decision boundaries between classes, are highly fragile to the number of pixels per class (*i.e.*, class imbalance), leading to decision boundaries that can be drastically altered by the majority classes (as shown in the left part of Figure 7). In contrast, diffusion-based generative models use log-likelihood as their optimization objective, maximizing the log-likelihood between the explicitly modeled class distribution and the underlying real one (as shown in the right part of Figure 7). Benefiting from modeling probabilistic density, diffusion-based generative models pay more attention to the class distribution rather than the boundaries across classes. Therefore, they conceptually exhibit *better tolerance to class imbalance*.

B PSEUDO ALGORITHM

In this section, we summarize the pseudo algorithm of DiffMatch in Algorithm 1. The inputs consist of a labeled set $\mathcal{D}^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N^l}$ and an unlabeled set $\mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$, where $N^u \gg N^l$. The feature extractor $g(\cdot)$, mask denoiser $f(\cdot)$, weak augmentation ω, strong augmentation s, and time step t are defined. The algorithm iterates over each batch of labeled data $(\boldsymbol{x}^l, \boldsymbol{y}^l)$ and unlabeled data $\boldsymbol{x}^u$. For labeled data, the pixel embedding $\boldsymbol{z}^l$ is extracted using $g(\cdot)$, noise is injected into $\boldsymbol{y}^l$ to obtain $\boldsymbol{y}_t^l$ via the forward process (Equation 3), and the noisy mask $\boldsymbol{y}_t^l$ is denoised conditioned on $\boldsymbol{z}^l$ and t using $f(\cdot)$ in the reverse process. The supervised loss $\mathcal{L}_{sup}$ is calculated by Equation 10. For unlabeled data, weak and strong augmentations are applied on $\boldsymbol{x}^u$ to obtain $\boldsymbol{x}_w^u$ and $\boldsymbol{x}_s^u$ respectively. Their pixel embeddings $\boldsymbol{z}_w^u$ and $\boldsymbol{z}_s^u$ are extracted using $g(\cdot)$. The pseudo label $\widetilde{\boldsymbol{y}}_{0,w}^u$ is obtained by denoising $\boldsymbol{\epsilon}$ conditioned on $\boldsymbol{z}_w^u$ in the reverse process. Noise is injected into $\widetilde{\boldsymbol{y}}_{0,w}^u$ to obtain $\widetilde{\boldsymbol{y}}_{t,w}^u$ via the forward process, and the noisy mask $\widetilde{\boldsymbol{y}}_{t,w}^u$ is denoised conditioned on $\boldsymbol{z}_s^u$ using $f(\cdot)$ in the reverse process. The unsupervised loss $\mathcal{L}_{unsup}$ is calculated by Equation 11. Finally, the model is updated by performing gradient backward on $\mathcal{L}_{sup} + \mathcal{L}_{unsup}$.

Algorithm 1 Pseudo algorithms of DiffMatch.

Inputs: Labeled Set $\mathcal{D}^l = \{(\boldsymbol{x}_i^l, \boldsymbol{y}_i^l)\}_{i=1}^{N^l}$, Unlabeled Set $\mathcal{D}^u = \{\boldsymbol{x}_i^u\}_{i=1}^{N^u}$ $(N^u \gg N^l)$
Define: Feature Extractor $g(\cdot)$, Mask Denoiser $f(\cdot)$, Weak Augmentation w, Strong Augmentation s, time step t
Output: Feature Extractor $g(\cdot)$, Mask Denoiser $f(\cdot)$
for each batch of $(\boldsymbol{x}^l, \boldsymbol{y}^l)$, $\boldsymbol{x}^u$ in $\mathcal{D}_l$, $\mathcal{D}_u$ **do**
    # Labeled Data:
    Extract pixel embedding $\boldsymbol{z}^l$ for $\boldsymbol{x}^l$ using $g(\cdot)$
    Inject noise into $\boldsymbol{y}^l$ and obtain $\boldsymbol{y}_t^l$ by Equation 3 ▷ *Forward Process*
    Denoise the noisy mask $\boldsymbol{y}_t^l$ conditioned on $\boldsymbol{z}^l$ and t using $f(\cdot)$ ▷ *Reverse Process*
    Calculate $\mathcal{L}_{sup}$ by Equation 10 ▷ *Supervised Loss*
    # Unlabeled Data:
    Obtain $\boldsymbol{x}_w^u$ and $\boldsymbol{x}_s^u$ by applying weak and strong augmentation on $\boldsymbol{x}^u$, respectively
    Extract pixel embedding $\boldsymbol{z}_w^u$ and $\boldsymbol{z}_s^u$ using $g(\cdot)$
    Obtain the pseudo label $\widetilde{\boldsymbol{y}}_{0,w}^u$ by denoising $\boldsymbol{\epsilon}$ conditioned on $\boldsymbol{z}_w^u$ using $f(\cdot)$
    ▷ *Reverse Process*
    Inject noise into $\widetilde{\boldsymbol{y}}_{0,w}^u$ and obtain $\widetilde{\boldsymbol{y}}_{t,w}^u$ by Equation 3 ▷ *Forward Process*
    Denoise the noisy mask $\widetilde{\boldsymbol{y}}_{t,w}^u$ conditioned on $\boldsymbol{z}_s^u$ using $f(\cdot)$ ▷ *Reverse Process*
    Calculate $\mathcal{L}_{unsup}$ by Equation 11 ▷ *Unsupervised Loss*
    Gradient backward $\mathcal{L}_{sup} + \mathcal{L}_{unsup}$ ▷ *Update Model*
end for

Algorithm 2 Diffusion Training Process

```
def alpha_cumprod(t, ns=0.0002, ds=0.00025):
  """cosine noise schedule"""
  n = torch.cos((t + ns) / (1 + ds) * math.pi / 2) ** -2
  return -torch.log(n - 1, eps=1e-5)

def train(images, masks):
  """images: [b, 3, h, w], masks: [b, 1, h, w]"""
  img_enc = feature_extractor(images) # encode image
  mask_enc = encoding(masks) # encode gt or pseudo labels
  mask_enc = (sigmoid(mask_enc) * 2 - 1) * scale # corrupt gt or pseudo
      labels
  eps = uniform(0, 1), normal(mean=0, std=1)
  mask_crpt = sqrt(alpha_cumprod(t)) * mask_enc + sqrt(1 - alpha_cumprod(t
      )) * eps
  # predict and backward
  mask_pred = mask_denoiser(mask_crpt, mask_enc, t)
  loss = l2_loss(mask_pred, masks) # calculate the loss after debiased
      adjustment
  return loss
```

Algorithm 3 Diffusion Sampling Process

```
def ddim(mask_t, mask_pred, t_now, t_next):
  """ estimate x at t_next with DDIM update rule"""
  α_now = alpha_cumprod(t_now)
  α_next = alpha_cumprod(t_next)
  mask_enc = encoding(mask_pred)
  mask_enc = (sigmoid(mask_enc) * 2 - 1) * scale
  eps = 1/√(1−α_now) * (mask_t − √α_now * mask_enc)
  mask_next = √α_next * x_pred + √(1 − α_now) * eps
  return mask_next

def sample(images, steps, td=1):
  """steps: sample steps, td: time difference"""
  img_enc = feature_extractor(images)
  mask_t = normal(0, 1) # [b, 256, h/4, w/4]
  for step in range(steps):
    # time intervals
    t_now = 1 - step / steps
    t_next = max(1 - (step + 1 + td) / steps, 0)
    # predict mask_0 from mask_t
    mask_pred = mask_denoiser(mask_t, img_enc, t_now)
    # estimate mask_t at t_next
    mask_t = ddim(mask_t, mask_pred, t_now, t_next)
  return mask_pred
```

C DERIVATION OF $\mathcal{L}^*_{diff}$

Here, we present the detailed derivation of $\mathcal{L}^*_{diff}$ from the learning of the diffusion model. Denoting the underlying conditional distribution as $\hat{q}$, we can rewrite the conditional reverse probability $\hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)$ according to Bayes' formula following Dhariwal & Nichol (2021):

$$\begin{aligned}\hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) =& \frac{q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}\right) \hat{q}\left(\boldsymbol{z} \mid \boldsymbol{y}_t\right)}{\hat{q}\left(\boldsymbol{z} \mid \boldsymbol{y}_{t+1}\right)} \\ =& \frac{q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}\right) \hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{z}\right) \hat{q}\left(\boldsymbol{z}\right) \hat{q}\left(\boldsymbol{y}_{t+1}\right)}{\hat{q}\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{z}\right) \hat{q}\left(\boldsymbol{z}\right) \hat{q}\left(\boldsymbol{y}_t\right)} \\ =& \frac{q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}\right) \hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{z}\right) \hat{q}\left(\boldsymbol{y}_{t+1}\right)}{\hat{q}\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{z}\right) \hat{q}\left(\boldsymbol{y}_t\right)}.\end{aligned} \tag{14}$$

Since the conditional diffusion model is trained to fit a prior distribution with known conditions by definition, we can approximate $\hat{q}(\boldsymbol{y}_t)$ with $p_\theta(\boldsymbol{y}_t)$ and have:

$$p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) = \frac{q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}\right) \hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{z}\right) \hat{q}\left(\boldsymbol{y}_{t+1}\right)}{\hat{q}\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{z}\right) p_\theta\left(\boldsymbol{y}_t\right)}. \tag{15}$$

Given the long tailed nature of the class distribution $p(\boldsymbol{y}_0)$ in practice, the learned conditional inverse probability $p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)$ is inevitably biased. To further reduce the risk of overfitting to the head classes and to increase coverage of the tail class distribution, we propose the debiased adjustment. First, we represent the conditional inverse probability under ideal condition (*i.e.*, when the class distribution is uniform, $p^*(\boldsymbol{y}_0) = \frac{1}{C}$, where C is the number of classes) as $p^*_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$. In the same way:

$$p^*_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) = \frac{q\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}\right) \hat{q}^*\left(\boldsymbol{y}_t \mid \boldsymbol{z}\right) \hat{q}^*\left(\boldsymbol{y}_{t+1}\right)}{\hat{q}^*\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{z}\right) p^*_\theta\left(\boldsymbol{y}_t\right)}. \tag{16}$$

Since $\boldsymbol{y}_0$ is uniquely determined by $\boldsymbol{z}$, we have:

$$\hat{q}^*\left(\boldsymbol{y}_t \mid \boldsymbol{z}\right) = \hat{q}^*\left(\boldsymbol{y}_t \mid \boldsymbol{y}_0\right) \overset{①}{=} \hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_0\right) = \hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{z}\right), \tag{17}$$

where the equality ① holds because $\hat{q}^*\left(\boldsymbol{y}_t \mid \boldsymbol{y}_0\right)/\hat{q}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_0\right)$ is conditioned on $\boldsymbol{y}_0$, *i.e.*, unrelated to $p(\boldsymbol{y}_0)$. In the same way:

$$\hat{q}^*\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{z}\right) = \hat{q}\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{z}\right). \tag{18}$$

In other words, $\hat{q}^*(\boldsymbol{y}_t \mid \boldsymbol{z})$ and $\hat{q}^*(\boldsymbol{y}_{t+1} \mid \boldsymbol{z})$ are not affected by the class distribution. Combining the above equations, we have:

$$p^*_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) = p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) \frac{p_\theta\left(\boldsymbol{y}_t\right)}{p^*_\theta\left(\boldsymbol{y}_t\right)} \frac{\hat{q}^*\left(\boldsymbol{y}_{t+1}\right)}{\hat{q}\left(\boldsymbol{y}_{t+1}\right)}. \tag{19}$$

It can be seen that there is only a factor of difference (*i.e.*, $\frac{p_\theta(\boldsymbol{y}_t)}{p^*_\theta(\boldsymbol{y}_t)} \frac{\hat{q}^*(\boldsymbol{y}_{t+1})}{\hat{q}(\boldsymbol{y}_{t+1})}$) between the ideal conditional inverse process $p^*_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)$ and the actual conditional inverse process $p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)$. However, the factor is difficult to obtain directly. Therefore, We convert it into the training loss and gradually remove this difference during training. Since $\frac{\hat{q}^*(\boldsymbol{y}_{t+1})}{\hat{q}(\boldsymbol{y}_{t+1})}$ is independent of the model parameters, it follows from Menon et al. (2020) that the sign should be reversed when converting the post-hoc adjustment factors into the training loss, giving us:

$$p^*_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) = p_\theta\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right) \frac{p^*_\theta\left(\boldsymbol{y}_t\right)}{p_\theta\left(\boldsymbol{y}_t\right)}. \tag{20}$$

Then we get the unbiased loss for the conditional diffusion model by replacing the $p_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$ in Equation 2 with $p^*_\theta(\boldsymbol{y}_0 \mid \boldsymbol{y}_t, \boldsymbol{z})$:

$$
\begin{aligned}
\mathcal{L}_{diff}^{*} &= \sum_{t} D_{\mathrm{KL}}\left[q\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t-1}\right) \| p_{\theta}^{*}\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)\right] \\
&= \sum_{t} \mathbb{E}_{q}\left[-\log \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)}{q\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t-1}\right)}\right] \\
&= \sum_{t}\left\{\mathbb{E}_{q}\left[-\log \frac{p_{\theta}\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)}{q\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t-1}\right)}\right]+\mathbb{E}_{q}\left[-\log \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{t}\right)}{p_{\theta}\left(\boldsymbol{y}_{t}\right)}\right]\right\} \\
&= \sum_{t}\left\{D_{\mathrm{KL}}\left[q\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t-1}\right) \| p_{\theta}\left(\boldsymbol{y}_{t} \mid \boldsymbol{y}_{t+1}, \boldsymbol{z}\right)\right]+\mathbb{E}_{q}\left[-\log \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{t}\right)}{p_{\theta}\left(\boldsymbol{y}_{t}\right)}\right]\right\}.
\end{aligned}
\tag{21}
$$

Focus on the second item of the above equation:

$$
\begin{aligned}
& \sum_{t} \mathbb{E}_{q}\left[-\log \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{t}\right)}{p_{\theta}\left(\boldsymbol{y}_{t}\right)}\right] \\
=& \sum_{t} \mathbb{E}_{q}\left\{-\log \mathbb{E}_{p_{\theta}}\left[\frac{p_{\theta}^{*}\left(\boldsymbol{y}_{0}\right) \prod_{t^{\prime}=1}^{t} p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}} \mid \boldsymbol{y}_{t^{\prime}-1}\right)}{p_{\theta}\left(\boldsymbol{y}_{0}\right) \prod_{t^{\prime}=1}^{t} p_{\theta}\left(\boldsymbol{y}_{t^{\prime}} \mid \boldsymbol{y}_{t^{\prime}-1}\right)}\right]\right\} \\
=& \sum_{t} \mathbb{E}_{q}\left\{-\log \mathbb{E}_{p_{\theta}}\left[\frac{p_{\theta}^{*}\left(\boldsymbol{y}_{0}\right) \prod_{t^{\prime}=1}^{t} p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right) \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}}\right)}{p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1}\right)}}{p_{\theta}\left(\boldsymbol{y}_{0}\right) \prod_{t^{\prime}=1}^{t} p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right) \frac{p_{\theta}\left(\boldsymbol{y}_{t^{\prime}}\right)}{p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1}\right)}}\right]\right\} \\
\overset{②}{\leqslant}& \sum_{t} \mathbb{E}_{q}\left\{\mathbb{E}_{p_{\theta}}\left[-\log \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{0}\right) \prod_{t^{\prime}=1}^{t} p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right) \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}}\right)}{p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1}\right)}}{p_{\theta}\left(\boldsymbol{y}_{0}\right) \prod_{t^{\prime}=1}^{t} p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right) \frac{p_{\theta}\left(\boldsymbol{y}_{t^{\prime}}\right)}{p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1}\right)}}\right]\right\} \\
=& \sum_{t} \mathbb{E}_{q}\left\{\mathbb{E}_{p_{\theta}}\left[\sum_{t^{\prime}=1}^{t}-\log \frac{p_{\theta}\left(\boldsymbol{y}_{0}\right) p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}{p_{\theta}^{*}\left(\boldsymbol{y}_{0}\right) p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}\right]\right\} \\
=& \mathbb{E}_{q}\left\{\mathbb{E}_{p_{\theta}}\left[\sum_{t} \sum_{t^{\prime}=1}^{t}-\log \frac{p_{\theta}\left(\boldsymbol{y}_{0}\right) p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}{p_{\theta}^{*}\left(\boldsymbol{y}_{0}\right) p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}\right]\right\} \\
\overset{③}{=}& \mathbb{E}_{q}\left[t \sum_{t}-\log \frac{p_{\theta}\left(\boldsymbol{y}_{0}\right) p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}{p_{\theta}^{*}\left(\boldsymbol{y}_{0}\right) p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}\right] \\
\overset{④}{=}& \sum_{t} t \mathbb{E}_{q}\left[-\log \frac{p_{\theta}^{*}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}{\frac{p_{\theta}\left(\boldsymbol{y}_{t^{\prime}-1} \mid \boldsymbol{y}_{t^{\prime}}\right)}{C p_{\theta}\left(\boldsymbol{y}_{0}\right)}}\right] \\
=& \sum_{t} t D_{\mathrm{KL}}\left[\frac{p_{\theta}\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_{t}\right)}{C p_{\theta}\left(\boldsymbol{y}_{0}\right)} \| p_{\theta}^{*}\left(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_{t}\right)\right],
\end{aligned}
\tag{22}
$$

where the inequality ② holds due to Jensen's Inequality, the equality ③ is valid because of the exchange in the order of summation, and the equality ④ is holds because $p^{*}(\boldsymbol{y}_0) = \frac{1}{C}$. In practice, we approximate the $p_\theta(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t)$ with Monte-Carlo samples from $p_\theta(\boldsymbol{y}_{t-1} \mid \boldsymbol{y}_t, \boldsymbol{z})$ and the loss reduce to:

$$
\begin{aligned}
\mathcal{L}_{diff}^{*} =& \|f\left(\boldsymbol{y}_{t}, \boldsymbol{z}\right)-\boldsymbol{y}_{0})\|^{2}+\tau t\left\|f\left(\boldsymbol{y}_{t}, \boldsymbol{z}\right)-\frac{f\left(\boldsymbol{y}_{t}, \boldsymbol{z}\right)}{C p\left(\boldsymbol{y}_{0}\right)}\right\|^{2} \\
=& \mathcal{L}_{diff}+\tau t\left\|f\left(\boldsymbol{y}_{t}, \boldsymbol{z}\right)-\frac{f\left(\boldsymbol{y}_{t}, \boldsymbol{z}\right)}{C p\left(\boldsymbol{y}_{0}\right)}\right\|^{2}.
\end{aligned}
\tag{23}
$$

D TRAINING CURVE FOR HEAD&TAIL CLASSES

Figure 8 provides a comparative analysis of the training samples of the head class *road* and the tail class *motorcycle* on the Cityscapes Cordts et al. (2016) under the 1/16 partition protocol as the training progresses. The proposed DiffMatch is compared with the highly competitive UniMatch Yang et al. (2022) in terms of pseudo label count, assuming that the ground truth for unlabeled data is available solely for theoretical analysis purposes.

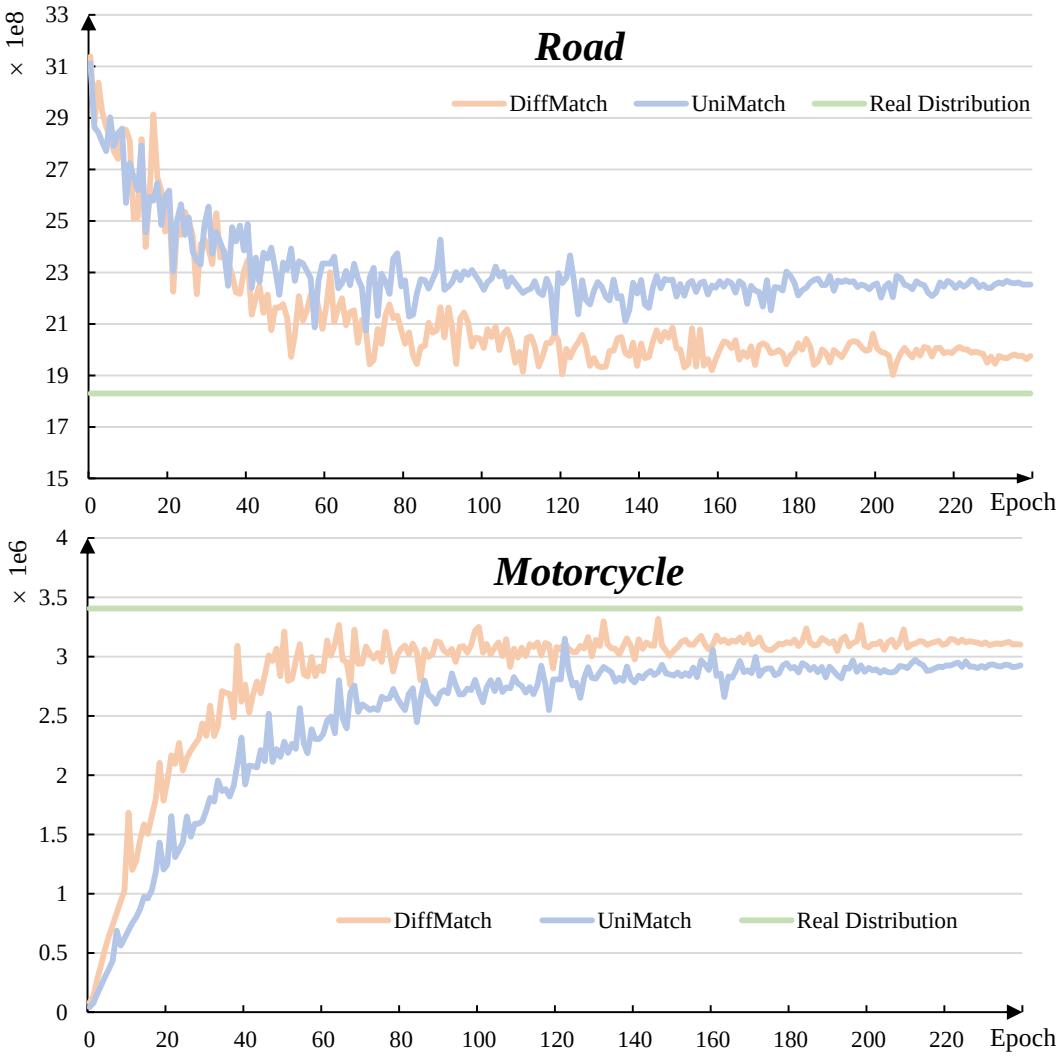


Figure 8: We count the training samples of **head class** *road* and **tail class** *motorcycle* on Cityscapes (Cordts et al., 2016) under 1/16 partition protocols as the training processes, and compare the proposed DiffMatch with the highly competitive UniMatch (Yang et al., 2022) in terms of *Pseudo Label Count*, assuming that the ground truth for unlabeled data is available solely for theoretical analysis purposes. Our DiffMatch strives to mitigate the Matthew effect raised by the class imbalance issue and stands out for head/tail classes.

The top plot in Figure 8 illustrates the prediction distribution of the head class *road*. UniMatch generates a significantly higher pseudo label count compared to the real distribution, indicating its tendency to over-predict the dominant class. In contrast, DiffMatch exhibits a pseudo label count that is more aligned with the real distribution, demonstrating its ability to mitigate the bias towards the head class. The bottom plot in Figure 8 depicts the prediction distribution of the tail class *motorcycle*. UniMatch generates substantially fewer pseudo labels compared to the real distribution, highlighting

its under-prediction of the minority class. Conversely, DiffMatch demonstrates a pseudo label count that is much closer to the real distribution, showcasing its effectiveness in addressing the under-representation of the tail class.

The comparative analysis in Figure 8 substantiates the effectiveness of DiffMatch in leveraging the advantages of generative models to alleviate the Matthew effect. By incorporating a diffusion model and theoretically deriving a debiased adjustment (3.3), DiffMatch effectively mitigates the bias towards head classes and the under-prediction of tail classes, promoting unbiased learning in semi-supervised semantic segmentation. This finding aligns with the quantitative results analyzed in the *"Performance in Head&Tail Classes"* of Section 4.3; please refer to it for more details.

E Detailed Analyses of Hyper-parameters

Decoder Depth. Table 7 investigates the effect of the decoder depth, *i.e.*, the number of layers in the mask denoiser $f(\cdot)$. The results demonstrate that increasing the depth initially improves the model accuracy, with the optimal performance achieved at 4 layers (73.3% mIoU on PASCAL *classic* 1/16(92)). However, further increasing the depth beyond 4 layers leads to the saturation of performance. This observation aligns with the goal behind the lightweight design of the mask denoiser, which enables efficient reuse of shared parameters during multi-step denoising diffusion processes while maintaining highly competitive performance. The chosen architecture with 4 layers strikes a balance between accuracy and efficiency, with a parameter count of 44.9M.

Table 7: Evaluation of number of layers

#Layer L	mIoU(92)	mIoU(1464)	#Param
1	70.1	79.5	42.4M
2	71.2	80.6	43.3M
4	**73.3**	**81.6**	44.9M
6	72.6	80.8	45.8M
12	71.9	81.1	49.9M

Scaling Factor. Table 8 explores the impact of the scaling factor b used in the analog bits encoding strategy (Section 3.4). The scaling factor determines the range $\{-b, b\}$ into which the analog bits are shifted and scaled. The results show that a suitable scaling factor is necessary for optimal performance. As the scaling factor increases, the model accuracy improves until reaching a peak at $b = 0.1$ (73.3% mIoU on PASCAL *classic* 1/16(92) and 81.6% mIoU on PASCAL *classic* Full(1464)). Further increasing the scaling factor leads to a decline in performance. We hypothesize that a larger scaling factor retains more easy cases with the same time step, potentially affecting the balance between easy and hard cases during training.

Table 8: Evaluation of scaling factor.

Scale b	mIoU(92)	mIoU(1464)
0.01	71.7	80.9
0.05	72.2	81.2
0.1	**73.3**	**81.6**
0.2	70.7	80.8
0.5	70.6	80.5

Regularization Term τ. Table 9 examines the influence of the trade-off weight τ for the regularization term in the debiased adjustment. The regularization term imposes a constraint between the prediction of mask denoiser and its roughly debiased version, reducing the risk of overfitting to head classes and increasing coverage of tail class distribution. The results indicate that setting $\tau = 0.1$ yields the optimal performance, that is, 73.3% mIoU on PASCAL *classic* 1/16(92) and 81.6% mIoU on PASCAL *classic* Full(1464).

Table 9: Evaluation of trade-off weight for the regularization term τ.

τ	mIoU(92)	mIoU(1464)
0.01	71.8	80.0
0.02	72.2	80.4
0.05	72.7	80.9
0.1	**73.3**	**81.6**
0.2	71.9	80.2

F Comparison with other Diffusion-based Semi-supervised Methods

As diffusion gains popularity in visual perception, researchers have introduced it into various semi-supervised tasks (You et al., 2024; Yang et al., 2024; Liu et al., 2024b; Ho et al., 2023), such as classification, federated learning, time-series classification and 3d object detection. In the following, we will comprehensively and meticulously compare our DiffMatch with these diffusion-based semi-supervised methods and summarize in Table 10 to highlight the originality of our work.

Different from our DiffMatch, both DPT (You et al., 2024) and FedDISC (Yang et al., 2024) *utilize an external diffusion model* to generate additional data and demonstrate their effectiveness in facilitating the original model training. Specifically, DPT introduces a from-scratch diffusion-based conditional generative model to address the scarcity of labeled data in semi-supervised classification task in three stages: train the original classifier on limited labeled data to predict pseudo-labels; train the conditional generative model using these pseudo-labels to generate labeled data; retrain the classifier with a combination of limited real and vast generated labeled data. FedDISC addresses the challenge of semi-supervised federated learning by introducing a well-trained diffusion model. To alleviate the communication burden between the server and clients, the diffusion model generates rich client-style data for the server, conditioned on the cluster centroid of client data representations, thereby facilitating model training on the server.

Regarding DiffShape (Liu et al., 2024b), although it explores integrating the diffusion process into semi-supervised time-series classification, it does so through a self-supervised mechanism *rather than incorporating it into the teacher-student network paradigm*. Specifically, DiffShape employs large amounts of unlabeled instance subsequences as conditions in the diffusion process to generate the subsequences themselves, enhancing similarity in the generated sequences compared to the original ones, thereby improving representation capability in a self-supervised manner.

For Diffusion-ss3d (Ho et al., 2023), although it integrates the diffusion process into the teacher-student network paradigm in semi-supervised 3D object detection, we categorize it as *a noise-to-filter paradigm*, leveraging the denoising capability of diffusion models to generate more accurate 3D bounding boxes as pseudo labels. Specifically, Diffusion-ss3d first predicts coarse bounding boxes (fixed bounding box candidate points) with a detection model, which can be considered as intermediate states in the diffusion process, and then employs the diffusion model as a denoising process to obtain other parameters of the bounding box (*e.g.*, bounding box size). Overall, this paradigm *partially exploits the characteristics of the diffusion process*, that is, the denoising ability, to improve the quality of the bounding boxes prediction.

Distinguished from these methods, Our DiffMatch integrates the diffusion process into the teacher-student network for semi-supervised semantic segmentation, which can be viewed as *a noise-to-prediction paradigm*. Motivated by the potential of generative models with better tolerance to class imbalance, our DiffMatch *learns the complete process* of transforming noise from a known distribution to class predictions (all states from time 0 to time T). Additionally, we *mathematically derive a debiased adjustment based on the state transition function* encapsulated in the diffusion process to further mitigate the Matthew effect. This mathematical formulation translates into strong empirical performance on real-world datasets, particularly in scenarios with the most limited labeled data and the most severe class imbalance. In general, DiffMatch completely utilizes the characteristics of the diffusion process in a different problem for semi-supervised semantic segmentation, aiming to provide a new perspective to alleviate the Matthew effect.

Table 10: Comparison with other diffusion-based semi-supervised methods.

	DPT	FedDISC	DiffShape	Diffusion-ss3d	**DiffMatch (Ours)**
Task	**classification**	**federated learning**	**time-series classification**	**3d object detection**	**semantic segmentation**
Motivation	harnessing the **data generation capability** of Diffusion to **alleviate data scarcity**	harnessing the **data generation capability** of Diffusion to **alleviate data scarcity**	using diffusion in a **self-supervised manner** to **improve representation capability**	exploiting the **denoising ability** of Diffusion to **improve the quality of pseudo label**	leveraging the **well class-imbalance tolerance** of Diffusion to **alleviate the Matthew effect**
Implementation	introducing a from-scratch **external diffusion model**	introducing a well-trained **external diffusion model**	integrating the diffusion process through a **self-supervised mechanism**	integrating the diffusion process into the teacher-student framework in a **noise-to-filter paradigm**	integrating the diffusion process into the teacher-student framework in a **noise-to-prediction paradigm**
Note				learning an **incomplete** diffusion process	(1) learning a **complete** diffusion process (2) **mathematically deriving a debiased adjustment** based on the state transition function

G Limitation and Society Impact

DiffMatch may face a potential limitation in terms of increased computational cost during multi-step inference. And how to adapt the number of inference steps to the degree of change in the generation state is a feasible direction. Within this paper, we present an approach for semi-supervised semantic segmentation, a pivotal research area in the realm of computer vision, with no apparent negative societal implications known thus far.

H Extended Discussion on Related Work

Semi-Supervised Segmentation. Semi-supervised semantic segmentation is a fundamental task with extensive applications in scene understanding (Mittal et al., 2019; Wu et al., 2023), medical image analysis (Yu et al., 2019; Bai et al., 2023; Zhao et al., 2025; Chi et al., 2024), and remote sensing interpretation (Wang et al., 2021; Bandara & Patel, 2022; Yuan et al., 2024). Owing to the recent advances in deep neural networks, semi-supervised semantic segmentation (Zhao et al., 2022b;a) has achieved conspicuous achievements. These algorithms leverage the mature combination of pseudo-labeling and consistency regularization (Lai et al., 2021; Zhong et al., 2021; Ouali et al., 2020; Chen & Lian, 2022) to improve performance. More recently, UniMatch (Yang et al., 2022) acknowledges the characteristics of semantic segmentation and incorporates appropriate data augmentations into FixMatch (Sohn et al., 2020), resulting in a concise yet powerful semi-supervised semantic segmentation baseline. Subsequently, a series of works aim to improve segmentation performance mainly in the following aspects. (1) Employ reasonable augmentation strategies to expand the augmentation space. For example, AugSeg (Zhao et al., 2023c) increases the randomness in RandAugment (Cubuk et al., 2020) for richer data augmentation space. iMAS (Zhao et al., 2023b) employs adaptive augmentations and supervisions conditioned on the model state. (2) Design effective teacher networks for better guidance. For example, Switch (Na et al., 2023) targets the coupling problem in the exponentially moving average (EMA) update process of teacher-student network and proposes a dual-teacher structure in an ensemble manner. (3) Utilize external knowledge to enhance the quality of pseudo labels. For example, LOGIC (Liang et al., 2023) integrates symbolic reasoning derived from symbolic knowledge to mitigate erroneous pseudo labels. SemiVL (Hoyer et al., 2025) incorporates a CLIP encoder (Radford et al., 2021), pre-trained on large-scale data, into semi-supervised semantic segmentation and employs a language-aware decoder to introduce text modality priors. (4) Enhance consistency regularization (Sun et al., 2024; Howlader et al., 2025b) to effectively exploit the information contained in unlabeled data. For example, RankMatch (Mai et al., 2024) utilizes inter-pixel correlations to construct more safe and effective supervision signals, which are in line with the nature of semantic segmentation. MPMC (Howlader et al., 2025a) identifies the classes present in an image region to incorporate pixel-level contextual information, thereby exploring more supervision signals. Despite yielding promising results, these methods tend to neglect the fact of class imbalance issue. In this paper, we strive to alleviate the negative impact (Matthew effect) raised by class imbalance issue and move towards unbiased semi-supervised learning.

Class-Imbalanced Semi-Supervised Segmentation. Real-world datasets usually yield a class-imbalanced distribution, especially in dense prediction tasks (*e.g.*, semantic segmentation), making the standard training of machine learning models harder to generalize. Existing methods to rebalance the training objective can be roughly categorized into two paradigms: (1) Re-sampling based methods (Chawla et al., 2002; He & Garcia, 2009; Byrd & Lipton, 2019; Chang et al., 2021; Shi et al., 2023; Wei et al., 2022) attempt to artificially balance the training data distribution. These approaches either employ over-sampling techniques to increase the representation of minority classes or utilize under-sampling strategies to reduce the dominance of majority classes. While effective in certain scenarios, these methods often struggle with the trade-off between maintaining data diversity and achieving balanced class distributions. (2) Re-weighting based methods (Cao et al., 2019; Cui et al., 2019; Huang et al., 2019; Ren et al., 2018; Hu et al., 2019; Chen et al., 2023d) focus on modifying the loss function to prioritize learning from under-represented classes. These approaches typically assign importance weights to different classes based on various criteria, such as inverse class frequency or dynamic class-wise difficulty measures. Although these methods have shown promising results, they often require careful tuning of weighting schemes to prevent instability during training. However, all these methods assume all labels are accessible to alleviate the class imbalance issue and thus inapplicable to the unlabelled data in semi-supervised semantic segmentation.

Recently, several studies have attempted to transfer these techniques on top of pseudo labels such as re-sampling (Wei et al., 2021), re-weighting (Wang et al., 2022a; Sun et al., 2023b; Xu et al., 2021; He et al., 2021; Wang et al., 2022b; Peng et al., 2023) (*e.g.*, Adsh (Guo & Li, 2022) utilizes adaptive thresholding that can be considered as binary weighting for semi-supervised learning, U^2PL (Wang et al., 2022b) adjusts the threshold adaptively to determine the reliability of pixels and constructs the extra supervised signal based on the negative classes of unreliable pixels, paying more attention to the tail classes), or a combination of both for semi-supervised learning (*e.g.*, AEL (Hu et al., 2021) adaptively balances the training of different categories). Nevertheless, these pseudo labels are often noisy as they are generated from poorly calibrated models. Furthermore, USRN (Guan et al., 2022) explores unbiased subclass regularization for alleviating the class imbalance issue. However, these discriminative methods are still confined to learning decision boundaries, which are brittle to the class imbalance issue, and the inherent nature of contempt for the underlying distribution remains unchanged. As a significant departure from the status quo, we formulate the semi-supervised semantic segmentation task as a conditional discrete data generation problem to model underlying distribution to overcome the shortcomings of discriminative solutions from a generative perspective.

Diffusion Models for Visual Perception. In addition to the significant progress in content generation, diffusion models have demonstrated incremental potential in the domain of perception (Chen et al., 2023b; Gu et al., 2022; Chen et al., 2023c; Brempong et al., 2022). Earlier studies primarily delve into investigating latent representations of diffusion models for zero-shot image segmentation (Baranchuk et al., 2021; Burgert et al., 2022) or applied diffusion models to medical image segmentation (Wolleb et al., 2022; Wu et al., 2022). Despite substantial progress, the outcomes of these efforts remain limited to specific local designs. The recent Pix2Seq-D (Chen et al., 2023c) extends the bit-diffusion (Chen et al., 2022) to panoptic segmentation, marking the first work of such expansion in a broader context. Additionally, DiffusionDet (Chen et al., 2023b) and DiffusionInst (Gu et al., 2022) explore diffusion models for query-based object detection (Carion et al., 2020) and instance segmentation (Zhang et al., 2021). Most recently, groundbreaking work has extended the application of diffusion models to a comprehensive range of dense visual perception tasks (Ji et al., 2023; Zhao et al., 2023a; Zheng et al., 2024). These latest developments have achieved promising results across multiple challenging scenarios, further solidifying the position of diffusion models as a versatile and powerful tool in the visual perception domain. Recently, several works have introduced diffusion into various semi-supervised tasks, such as classification, federated learning, time-series classification, and 3d object detection. Among them, both DPT (You et al., 2024) and FedDISC (Yang et al., 2024) aim to introduce an external diffusion model to generate data and utilize these data in a multi-stage training manner. DiffShape (Liu et al., 2024b) utilizes diffusion in a self-supervised manner to improve representation capability, and Diffusion-ss3d (Ho et al., 2023) exploits the denoising ability of the diffusion to improve the quality of the pseudo label. However, these methods differ from ours both from motivation to implementation. We comprehensively and meticulously compare our DiffMatch with these diffusion-based semi-supervised methods in Appendix F. In general, DiffMatch completely utilizes the characteristics of the diffusion process for semi-supervised semantic segmentation, aiming to provide a new perspective to alleviate the Matthew effect.

I More Visualization

Here, we provide additional visualizations to qualitatively assess the performance of DiffMatch in comparison to other methods. Figure 9 showcases the segmentation results on the PASCAL VOC dataset, highlighting the effectiveness of DiffMatch in obtaining more accurate semantic segmentation, particularly for pixels that are incorrectly segmented as the most dominant class by other methods. For example, in the 2nd row, FreeMatch, UniMatch, and RankMatch encounter difficulties in accurately segmenting the *person* pixels. They misclassify a considerable portion of the *person* pixels as the *horse* class. These misclassifications can be attributed to the class imbalance issue, where the models are inclined to favor the majority classes, resulting in subpar segmentation performance for the less represented classes like *person*. In contrast, DiffMatch demonstrates a notable ability to overcome these challenges and generate more precise segmentations. By incorporating a generative perspective and employing a debiased adjustment, DiffMatch effectively mitigates the Matthew effect stemming from class imbalance. As a result, it accurately segments the *person* pixels.

Furthermore, Figure 10 offers additional insights into the inference trajectory of DiffMatch across different diffusion sampling steps. The ground truth segmentation is provided as a reference, and the segmentation results at steps 1, 2, and 3 are visualized. As the number of sampling steps increases, the segmentation quality progressively improves, with finer details and more accurate boundary delineation.

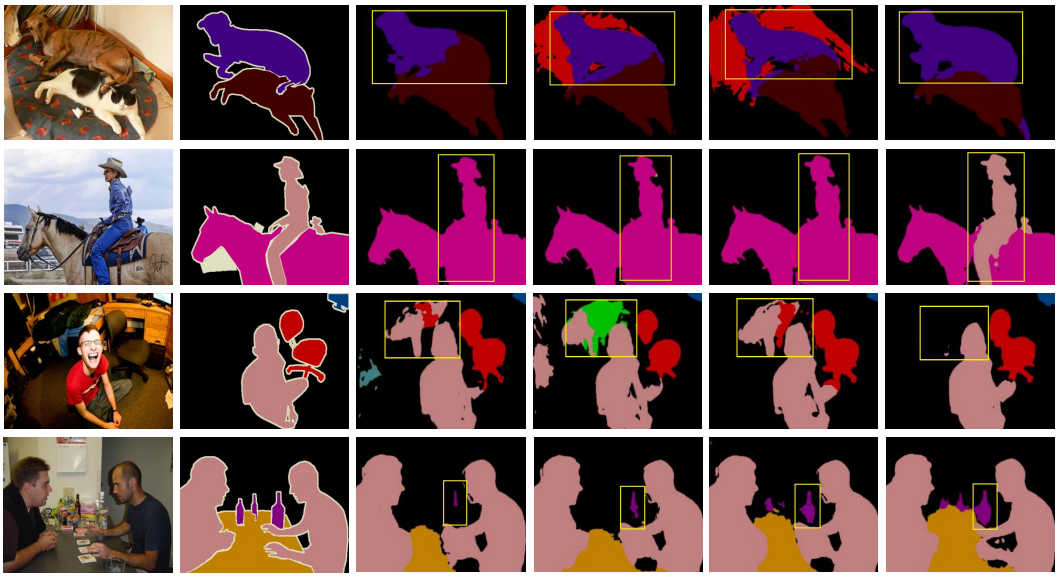


Figure 9: Qualitative results on PASCAL VOC dataset. DiffMatch can obtain more accurate segmentation for pixels that are inaccurately segmented as the most dominant class.

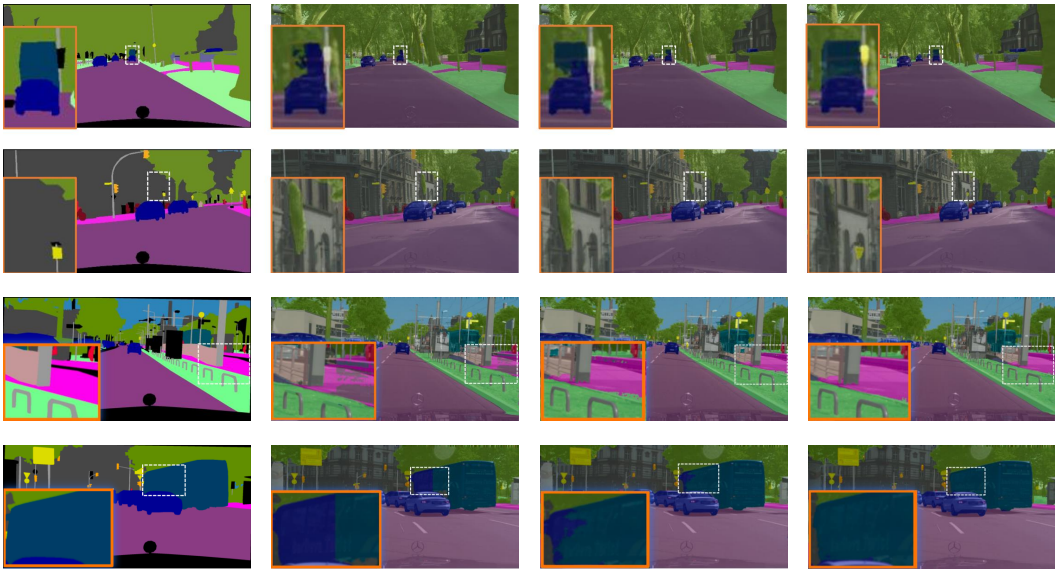


Figure 10: Inference trajectory with diffusion sampling steps. The model gradually refines the prediction, starting from a coarse estimation in Step 1 and progressively improving the results in Step 2. The final output in Step 3 closely resembles the ground truth, demonstrating the effectiveness of DiffMatch in capturing fine-grained details and accurately delineating the boundaries of the changed buildings.

J SCALABILITY FOR OTHER SCENARIOS

To further demonstrate the versatility and practicality of DiffMatch, we extend our experiments beyond natural image benchmarks and explore its performance in two crucial real-world applications: remote sensing interpretation and medical image analysis. These domains are characterized by an abundance of unlabeled data but a scarcity of manual annotations due to the high cost of expert labeling.

J.1 REMOTE SENSING INTERPRETATION SCENARIO

Remote Sensing Interpretation. We further conduct extra experiments on the widely used change detection dataset WHU-CD (Bandara & Patel, 2022; Liu et al., 2020) to evaluate the scalability of our method. The WHU-CD dataset, primarily designed for detecting changes in buildings, is particularly challenging due to its large change area and highly skewed class distribution. As illustrated in Table 11, **the pixel count of the head class *unchanged* is over 21 times that of the tail class *changed***, posing a significant challenge for the change detection task due to the extreme class imbalance issue. To thoroughly evaluate the effectiveness of DiffMatch, we split the WHU-CD dataset into three subsets following previous methods (Yang et al., 2022): a training set containing 5,947 images, a verification set with 743 images, and a test set comprising 744 images.

Table 11: Class distribution statistics on WHU-CD dataset.

Class Name	Unchanged	Changed
Ratio	95.55%	4.45%

Table 12 presents the quantitative results of various SSL methods (S4GAN (Bandara & Patel, 2022), SemiCDNet (Mittal et al., 2019), SemiCD (Peng et al., 2020), UniMatch (Yang et al., 2022)) on the WHU-CD dataset under different partition protocols. DiffMatch consistently outperforms all other methods across all labeled data ratios, with the performance gap being most significant when the labeled data is scarce. Specifically, with only 5% labeled data, DiffMatch achieves an changed-class IoU of 80.7%, surpassing the supervised baseline by a remarkable 32.4% and outperforming the second-best method, UniMatch, by 3.2%. These results demonstrate the robustness and effectiveness of DiffMatch in tackling the challenging change detection task, especially in low-data regimes where the class imbalance issue is most severe. The superior performance of DiffMatch can be attributed to its ability to effectively mitigate the Matthew effect through its generative modeling approach and debiased adjustment strategy.

Table 12: Quantitative results of different SSL methods on WHU-CD dataset. We report changed-class IOU (%) under various partition protocols and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Method	PSPNet			
	5%	10%	20%	40%
Sup.-only	48.3	60.7	69.7	69.5
S4GAN	18.3	62.2	70.8	76.4
SemiCDNet	51.7	62.0	66.7	75.9
SemiCD	65.8	68.1	74.8	77.2
UniMatch	77.5	78.9	82.9	84.4
DiffMatch (Ours)	**80.7**	**81.6**	**84.8**	**86.3**
$\Delta\uparrow$	+32.4	+20.9	+15.1	+16.8

Table 13 summarizes that DiffMatch achieves the best results across all classes, showcasing its ability to handle the class imbalance issue effectively. This significant improvement demonstrates DiffMatch's effectiveness in enhancing the performance of the underrepresented class, which is often challenging for discriminative models due to the scarcity of labeled data.

The qualitative results in Figure 11 show that DiffMatch generates predictions that closely resemble the ground truth, accurately detecting the changed building areas with precise boundaries. In contrast, discriminative learning-based methods like UniMatch and SemiCD are more affected by the class imbalance issue, incorrectly classifying background pixels as changed, resulting in noticeable noise in their predictions. This further validates the effectiveness of DiffMatch in alleviating the class imbalance problem from a generative perspective.

Table 13: Per-class performance comparison of different methods on WHU-CD dataset under the 5% labeled data setting. The **best** is highlighted in **bold**.

Method	Unchanged (head)	Changed (tail)
Sup.-only	93.2	48.3
SemiCD	96.3	65.8
UniMatch	97.9	77.5
DiffMatch (Ours)	**98.5**	**80.7**

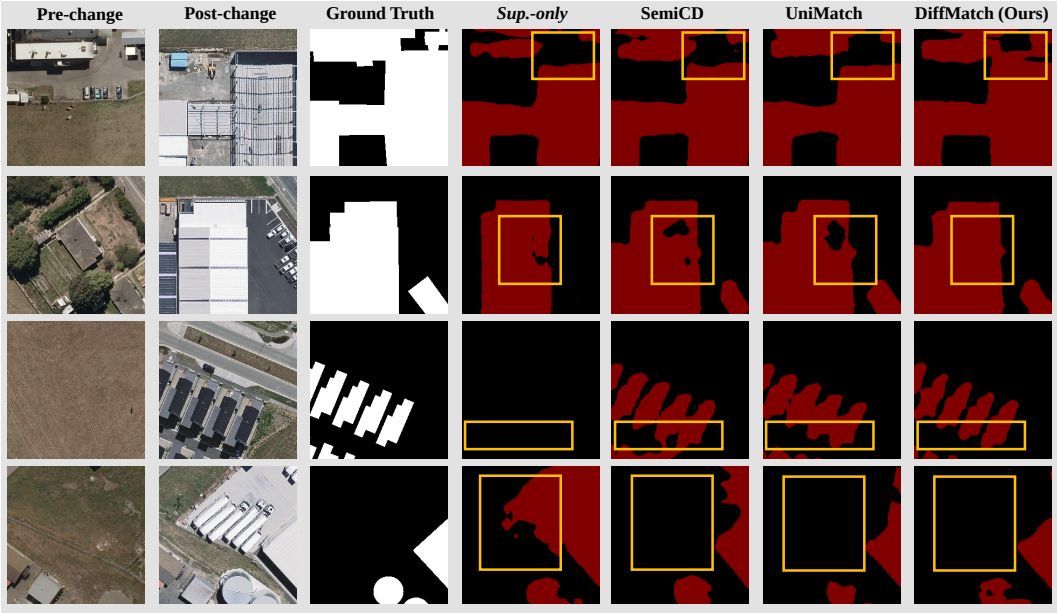


Figure 11: Qualitative results on WHU-CD dataset.

J.2 Medical Image Analysis Scenario

Medical Image Analysis–ACDC. To assess the applicability of our method in the medical domain, we conduct experiments on the ACDC (Bernard et al., 2018) dataset for semi-supervised cardiac MRI segmentation. The ACDC dataset consists of 100 patient scans, each with manual annotations of the background, right ventricle (RV), myocardium (MYO), and left ventricle (LV). The dataset poses challenges due to the limited number of annotated samples and the inherent class imbalance among the target structures. Table 14 illustrates the highly skewed pixel-level class distribution, with the head class *background* dominating at 96.20% and the tail class *right ventricle* (RV) constituting a mere 1.18%, **with over 81 head-to-tail ratio**. Following the standard protocol (Yang et al., 2022), we split the ACDC dataset into three subsets: 70 scans for training, 10 scans for validation, and 20 scans for testing. We evaluate the effectiveness of our proposed DiffMatch under different labeled data ratios (*i.e.*, 1, 3, and 7 labeled cases) to simulate real-world scenarios where expert annotations are scarce and expensive to obtain.

Table 15 showcases the quantitative results of various SSL methods (UniMatchCNN (Yang et al., 2022), &Trans (Luo et al., 2022), CPS (Chen et al., 2021b), UA-MT (Yu et al., 2019)) on the ACDC

Table 14: Class distribution statistics on ACDC dataset.

Class Name	Background	Myocardium	Left Ventricle	Right Ventricle
Ratio	96.20%	1.33%	1.29%	1.18%

dataset, reporting the mean Dice Similarity Coefficient (DSC) for the class of interest (RV, MYO, LV) under different labeled data amounts. DiffMatch consistently achieves the best performance across all settings, with significant improvements over the supervised baseline (*Sup.-only*). Notably, with only a single labeled case, DiffMatch obtains a DSC of 87.3%, outperforming the supervised baseline by a substantial margin of 58.8%. As the number of labeled cases increases to 3 and 7, DiffMatch maintains its lead over UniMatch, with performance gaps of 1.6% and 1.1%, respectively. These results demonstrate the robustness and scalability of DiffMatch in handling various levels of data scarcity, consistently outperforming UniMatch and other SSL methods. Overall, the results on the ACDC dataset validate the applicability and effectiveness of DiffMatch in the medical domain, showcasing its potential to alleviate the reliance on extensive expert annotations and improve segmentation performance in semi-supervised settings.

Table 15: Quantitative results of different SSL methods on ACDC dataset. We report mean Dice Similarity Coefficient (DSC) (%) with various labeled cases and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Method	UNet		
	1 case	3 cases	7 cases
Sup.-only	28.5	41.5	62.5
UA-MT	-	61.0	81.5
CPS	-	60.3	83.3
CNN&Trans	-	65.6	86.4
UniMatch	85.4	88.9	89.9
DiffMatch (Ours)	**87.3**	**90.5**	**91.0**
$\Delta\uparrow$	+58.8	+49.0	+28.5

Table 16 presents a comprehensive comparison of DiffMatch against state-of-the-art semi-supervised methods on the ACDC dataset for the class of interest, evaluating the Dice coefficient for each class. DiffMatch consistently achieves the best performance across all classes, demonstrating substantial improvements over the supervised baseline, especially in the extreme low-data regime with only a single labeled case. Notably, DiffMatch exhibits robust performance even for the most underrepresented RV class, underscoring the merits of generative modeling in tackling class imbalance.

Table 16: Per-class performance comparison of different methods for the class of interest (MYO, LV, RV) on ACDC dataset under the 3 labeled case setting. The **best** is highlighted in **bold**.

Method	Myocardium	Left Ventricle	Right Ventricle
Sup.-only	43.7	52.1	28.7
CPS	65.2	72.0	43.8
UniMatch	89.3	98.2	78.6
DiffMatch (Ours)	**91.5**	**99.3**	**80.7**

The qualitative results in Figure 12 showcase the superiority of DiffMatch in generating accurate and coherent segmentations for the classes of interest. Compared to the supervised baseline (Sup.-only) and other SSL methods, DiffMatch produces segmentations that closely resemble the ground truth, capturing fine details and maintaining precise boundaries around the heart. Notably, for the challenging myocardium (MYO) class, DiffMatch demonstrates a remarkable ability to segment this

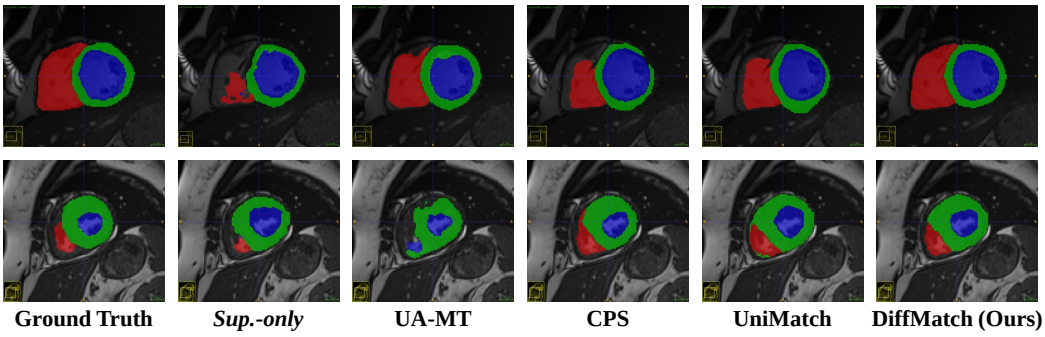


Figure 12: Qualitative results on ACDC dataset.

thin structure accurately, while other methods either over-segment or miss portions of the MYO. This visual comparison further validates DiffMatch's effectiveness in handling class imbalance and its applicability in the medical domain, where accurate segmentation of underrepresented structures is crucial for diagnosis and treatment planning.

Medical Image Analysis–Synapse. Furthermore, we extend our experiments to the more challenging Synapse (Tang et al., 2022; Igelsias et al., 2015) dataset to validate the scalability of our method. Synapse is a multi-organ segmentation dataset that contains 30 axial abdominal CT scan cases (3,779) with 8 manually annotated abdominal organ classes (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach). Compared to the ACDC dataset, Synapse exhibits more severe class imbalance and anatomical variability. Table 17 summarizes the class distribution statistics on the Synapse dataset, revealing a severe class imbalance among the abdominal organs. **According to statistics, the ratio of head class *liver* to tail class *left kidney* reaches 593.** This extreme imbalance poses significant challenges for accurate multi-organ segmentation. To fairly evaluate the effectiveness of DiffMatch (Chen et al., 2021a; Liu et al., 2024a), we split Synapse into 18 cases for training (2,212 slices) and 12 cases for testing. We evaluate performance on the Synapse dataset under different labeled data ratios (*i.e.*, 4, 2, and 1 labeled cases) to simulate real-world scenarios where expert annotations are extremely scarce and expensive to obtain.

Table 17: Class distribution statistics on Synapse dataset.

Class Name	Liver	Stom	Pancr	Spleen	Aorta	Gallb	Kid_R	Kid_L
Ratio	71.23%	15.20%	6.20%	3.91%	2.61%	0.56%	0.18%	0.12%

Table 18: Quantitative results of different SSL methods on Synapse dataset. We report mean Dice Similarity Coefficient (DSC) (%) with various labeled cases and show the improvements over *Sup.-only* baseline. The **best** is highlighted in **bold**.

Method	UNet		
	1 case	2 cases	4 cases
Sup.-only	10.7	42.5	51.9
CPS	15.0	48.8	57.9
CTS	26.3	55.2	64.0
MCSC	34.0	61.1	68.5
UniMatch	41.1	64.0	69.3
DiffMatch (Ours)	**44.3**	**66.1**	**70.6**
$\Delta\uparrow$	+33.6	+23.6	+18.7

Table 18 showcases the quantitative results of various SSL methods (CTS (Luo et al., 2022), CPS (Chen et al., 2021b), MCSC (Liu et al., 2023), UniMatch (Yang et al., 2022)) on the Synapse

dataset, reporting the mean Dice Similarity Coefficient (DSC) for eight classes (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach) under different labeled data amounts. DiffMatch consistently achieves the best performance across all settings, with significant improvements over the supervised baseline (Sup.-only). Notably, with only a single labeled case, DiffMatch obtains a DSC of 44.3%, outperforming the supervised baseline by a substantial margin of 33.6%. This demonstrates DiffMatch's effectiveness in leveraging unlabeled data to improve segmentation performance, even in extremely low-data regimes. As the number of labeled cases increases to 2 and 4, DiffMatch maintains its superior performance compared to other SSL methods. The superior performance of DiffMatch on the Synapse dataset highlights its effectiveness in handling severe class imbalance and limited annotations in multi-organ segmentation tasks.

Table 19: Per-class performance comparison of different methods (aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach) on Synapse dataset under the 1 labeled case setting. The **best** is highlighted in **bold**.

Method	Liver	Stom	Pancr	Spleen	Aorta	Gallb	Kid_R	Kid_L
Sup.-only	33.8	6.9	1.5	8.7	14.7	9.1	5.6	5.3
CPS	59.4	7.2	2.3	9.4	19.6	9.6	6.9	5.6
UniMatch	76.0	21.2	8.6	64.1	62.5	11.7	69.9	14.8
DiffMatch (Ours)	**78.4**	**27.7**	**12.7**	**66.8**	**63.8**	**14.2**	**72.1**	**18.7**

Table 19 provides a comprehensive comparison of the per-class performance of different methods on the Synapse dataset under the challenging 1 labeled case setting. DiffMatch consistently outperforms other methods across all classes, demonstrating its effectiveness in handling class imbalance. Notably, DiffMatch achieves substantial improvements over the supervised baseline (Sup.-only) for the minority classes, such as gallbladder (+5.1%), right kidney (+66.5%), and left kidney (+13.4%). These results highlight DiffMatch's ability to reduce the risk of overfitting to the head classes and increase coverage of the tail class distribution from a generative perspective. Moreover, DiffMatch maintains its superior performance for the majority classes, such as liver (+44.6%), indicating its robustness against class imbalance and limited annotations simultaneously.

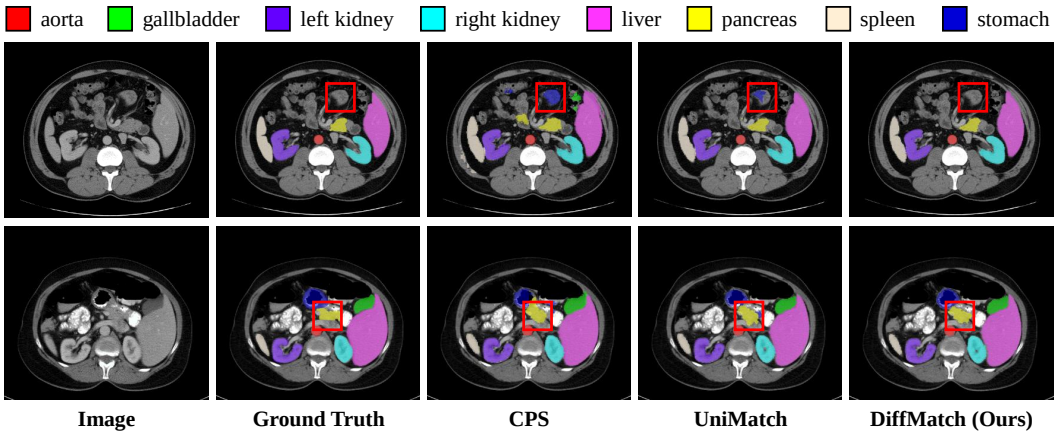


Figure 13: Qualitative results on Synapse dataset.

Figure 13 illustrates the qualitative results of different methods on the Synapse dataset, visually comparing the segmentation quality for various abdominal organs. We can observe that DiffMatch generates segmentations that closely resemble the ground truth, accurately delineating the boundaries of the organs and capturing fine-grained details. Notably, DiffMatch effectively captures the challenging minority classes, such as gallbladder and kidneys, which are often missed or poorly segmented by other methods. These qualitative results further validate DiffMatch's ability to handle severe class imbalance and limited annotations, as it consistently produces more accurate and coherent segmentations across all classes.

K A CLOSER LOOK AT DIFFMATCH FROM AN INFORMATION-THEORETIC PERSPECTIVE

Semi-supervised semantic segmentation is essentially a learning strategy-centric task, and its core lies in effectively exploring numerous unlabeled data under extremely limited labeled data. In this work, we aim to calibrate the model's biased predictions caused by the inherent class imbalance from a generative perspective and propose DiffMatch to integrate the complete diffusion process strategy into the teacher-student network for semi-supervised semantic segmentation.

From the perspective of constructing supervision signals to take a closer look, DiffMatch also harvests additional merits beyond regular consistency regularization learning, *i.e.*, the supervision signals constructed during the noise addition (forward diffusion transition) and denoising (reverse diffusion process) stages. This type of supervision signal inherits the spirit from the self-supervised learning community, known as *predictive learning*, an effective paradigm for mining information from unlabeled data (Gidaris et al., 2018; Lee et al., 2021; Doersch et al., 2015). It can be abstracted and summarized as learning to predict self-generated surrogate transformations, *i.e.*, by applying a transformation (*e.g.*, adding Gaussian noise) to the input and learning in a way that predicts the pattern of the transformation (*e.g.*, denoising). In this context, following the information-theoretic framework developed by (Tsai et al., 2021), we show that DiffMatch provably enjoys better pseudo-label quality.

We denote the random variable of the input image as $\mathbf{X}$ and get the pseudo-label prediction $\mathbf{Y}$ through the segmentation model $\mathcal{F}_{\boldsymbol{\theta}} : \mathbf{Y} = \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{X})$. From an information-theoretic learning perspective, a desirable strategy should maximize the Mutual Information (MI) between $\mathbf{Y}$ and $\mathbf{T}$, *i.e.*, $I(\mathbf{Y};\mathbf{T})$, where $\mathbf{T}$ is the assumed ground truth for unlabeled data[1]. Ideally, assuming we have access to the ground truth for all pseudo-labels, in this case, we can fully explore the information in the unlabeled data by directly maximizing $I(\mathbf{Y};\mathbf{T})$, and semi-supervised learning would be equivalent to fully supervised learning, thus yielding the performance upper bound for semi-supervised learning (*i.e.*, oracle performance).

However, in practice, without access to the ground truth $\mathbf{T}$, semi-supervised semantic segmentation instead resorts to maximizing $I(\mathbf{Y};\mathbf{S})$ by constructing effective surrogate supervision signals on unlabeled data, where $\mathbf{S}$ denotes the surrogate signals. In specific, consistency regularization learning aims to match $\mathbf{Y}$ with the student network's prediction based on the augmented view of the image, denoted as $\mathbf{S}_{cr}$; while $\mathbf{S}_{pr}$, derived from the predictive learning paradigm, seeks to use $\mathbf{Y}$ to predict the applied transformation (noise) guided by the image. In DiffMatch, since we enjoy the combination of these two surrogate supervision signals, we actually maximize the MI with respect to their joint distribution $I(\mathbf{Y};\mathbf{S}_{cr},\mathbf{S}_{pr})$. We denote the pseudo label predictions in the case of fully supervised learning, consistency regularization, and DiffMatch as $\mathbf{Y}_{oracle}$, $\mathbf{Y}_{cr}$, and $\mathbf{Y}_{diff}$, respectively.

We have the following inequalities (Theorem K.3) when the segmentation model $\mathcal{F}_{\boldsymbol{\theta}}$ is in the sufficient and minimal learning status, that is, $I\left(\mathbf{Y}_{oracle};\mathbf{T}\right) \geq I\left(\mathbf{Y}_{diff};\mathbf{T}\right) \geq I\left(\mathbf{Y}_{cr};\mathbf{T}\right)$, **indicating that DiffMatch theoretically improves the quality of pseudo labels** (manifested in the greater mutual information between the pseudo prediction $\mathbf{Y}_{diff}$ and the ground truth $\mathbf{T}$).

Below, we provide a complete proof for Theorem K.3. More rigorously, for a model with enough capacity, we give the definition of the sufficient and minimal learning status based on the surrogate supervision signals constructed from the unlabeled data (Tsai et al., 2021).

Definition K.1. A model with enough capacity is in the sufficient and minimal learning status for the surrogate supervision signal $\mathbf{S}$ if its pseudo label prediction $\mathbf{Y}^*$ satisfies the following conditions meantime: (1) the model's learning status is sufficient, when $\mathbf{Y}^* = \arg\max_{\mathbf{Y}} I\left(\mathbf{Y};\mathbf{S}\right)$; (2) the model's learning status is minimal, when $\mathbf{Y}^* = \arg\min_{\mathbf{Y}} H\left(\mathbf{Y} \mid \mathbf{S}\right)$.

Definition K.1 indicates that a model with enough capacity: (1) when in a sufficient learning status, the prediction $\mathbf{Y}^*$ can reflect as much information as possible contained in the surrogate supervision signal $\mathbf{S}$; (2) When in a minimal learning status, the prediction $\mathbf{Y}^*$ can reflect the information from the surrogate supervision signal $\mathbf{S}$ while minimizing redundancy. This is in contrast to underfitting

[1] Assuming that the ground truth for unlabeled data is available solely for theoretical analysis purposes.

(with insufficient capability to fully capture the surrogate supervision signal) and overfitting (sensitive to redundant information from augmented views or transformation patterns of the input) caused by insufficient or excessive model capacity, respectively.

Then we can derive the Lemma K.2 showing that the maximal mutual information of $I(\mathbf{Y}^*;\mathbf{S})$ is $I(\mathbf{X};\mathbf{S})$.

Lemma K.2. *For the pseudo label prediction $\mathbf{Y}$ obtained by the segmentation model $\mathcal{F}_{\boldsymbol{\theta}}$ with enough capacity in the sufficient and minimal learning status from the input image $\mathbf{X}$, we have $I(\mathbf{Y}^*;\mathbf{S}) = I(\mathbf{X};\mathbf{S})$.*

Proof. In fact, the pseudo label prediction $\mathbf{Y}$ is conditionally independent with surrogate supervision signal $\mathbf{S}$, given the input image $\mathbf{X}$, *i.e.*, $\mathbf{Y} \perp \mathbf{S} \mid \mathbf{X}$. Then, we can calculate as follows:

$$I(\mathbf{Y};\mathbf{S}) - I(\mathbf{X};\mathbf{S}) = [I(\mathbf{S};[\mathbf{Y},\mathbf{X}]) - I(\mathbf{X};\mathbf{S})] - [I(\mathbf{S};[\mathbf{Y},\mathbf{X}]) - I(\mathbf{Y};\mathbf{S})]. \tag{24}$$

Considering $I(\mathbf{S};[\mathbf{Y},\mathbf{X}]) = I(\mathbf{X};\mathbf{S}) + I(\mathbf{Y};\mathbf{S} \mid \mathbf{X})$, and $I(\mathbf{S};[\mathbf{Y},\mathbf{X}]) = I(\mathbf{Y};\mathbf{S}) + I(\mathbf{X};\mathbf{S} \mid \mathbf{Y})$, then Equation 24 can be formalized as,

$$\begin{aligned} I(\mathbf{Y};\mathbf{S}) - I(\mathbf{X};\mathbf{S}) &= [I(\mathbf{X};\mathbf{S}) + I(\mathbf{Y};\mathbf{S} \mid \mathbf{X}) - I(\mathbf{X};\mathbf{S})] - [I(\mathbf{Y};\mathbf{S}) + I(\mathbf{X};\mathbf{S} \mid \mathbf{Y}) - I(\mathbf{Y};\mathbf{S})] \\ &= I(\mathbf{Y};\mathbf{S} \mid \mathbf{X}) - I(\mathbf{X};\mathbf{S} \mid \mathbf{Y}). \end{aligned} \tag{25}$$

Since $\mathbf{Y} \perp \mathbf{S} \mid \mathbf{X}$, we have $I(\mathbf{Y};\mathbf{S} \mid \mathbf{X}) = 0$. Therefore, $I(\mathbf{Y};\mathbf{S}) - I(\mathbf{X};\mathbf{S}) = -I(\mathbf{X};\mathbf{S} \mid \mathbf{Y}) \leq 0$, that is, $I(\mathbf{Y};\mathbf{S}) \leq I(\mathbf{X};\mathbf{S})$. Then for the segmentation model $\mathcal{F}_{\boldsymbol{\theta}}$ with enough capacity in the sufficient and minimal learning status, the pseudo label prediction $\mathbf{Y}^*$ will have $I(\mathbf{Y}^*;\mathbf{S}) = \arg\max_{\mathbf{Y}} I(\mathbf{Y};\mathbf{S}) = I(\mathbf{X};\mathbf{S})$. □

Next, we derive Theorem K.3 based on the above definitions and lemma and provide a complete proof.

Theorem K.3. *We have the following inequalities when the segmentation model $\mathcal{F}_{\boldsymbol{\theta}}$ is in the sufficient and minimal learning status, $\mathbf{Y}_{oracle}$, $\mathbf{Y}_{cr}$, $\mathbf{Y}_{diff}$:*

$$I(\mathbf{Y}_{oracle};\mathbf{T}) \geq I(\mathbf{Y}_{diff};\mathbf{T}) \geq I(\mathbf{Y}_{cr};\mathbf{T}). \tag{26}$$

Proof. According to the Lemma K.2, we have the following properties for pseudo label prediction:

$$I(\mathbf{Y}_{cr};\mathbf{S}_{cr}) = I(\mathbf{X};\mathbf{S}_{cr}), I(\mathbf{Y}_{diff};\mathbf{S}_{cr},\mathbf{S}_{pr}) = I(\mathbf{X};\mathbf{S}_{cr},\mathbf{S}_{pr}). \tag{27}$$

Therefore, for the pseudo label prediction $\mathbf{Y} \in \{\mathbf{Y}_{cr}, \mathbf{Y}_{diff}\}$ obtained by the segmentation model with enough capacity in the sufficient and minimal learning status, and the corresponding surrogate supervision signal $\mathbf{S} \in \{\mathbf{S}_{cr}, (\mathbf{S}_{cr},\mathbf{S}_{pr})\}$, we have,

$$I(\mathbf{Y};\mathbf{S};\mathbf{T}) = I(\mathbf{X};\mathbf{S};\mathbf{T}), I(\mathbf{Y};\mathbf{S} \mid \mathbf{T}) = I(\mathbf{X};\mathbf{S} \mid \mathbf{T}). \tag{28}$$

Besides, because the segmentation model is in the sufficient and minimal learning status, we also have,

$$I(\mathbf{Y};\mathbf{T} \mid \mathbf{S}) \leq I(\mathbf{Y} \mid \mathbf{S}) = 0. \tag{29}$$

Together with the two equalities above, we further have the following equality on $I(\mathbf{Y};\mathbf{T})$:

$$\begin{aligned} I(\mathbf{Y};\mathbf{T}) &= I(\mathbf{Y};\mathbf{T};\mathbf{S}) + I(\mathbf{Y};\mathbf{T} \mid \mathbf{S}) \\ &= I(\mathbf{X};\mathbf{T};\mathbf{S}) + \underbrace{I(\mathbf{Y};\mathbf{T} \mid \mathbf{S})}_{0} \\ &= \underbrace{I(\mathbf{X};\mathbf{T})}_{\text{unchanged}} - I(\mathbf{X};\mathbf{T} \mid \mathbf{S}). \end{aligned} \tag{30}$$

For the fully supervised learning (oracle), considering that it is an ideal condition where all the ground truth $\mathbf{T}$ corresponding to the unlabeled data is accessible for training based on the supervision signal $\mathbf{S}_{oracle}$, that is, all the pseudo labels are correct, in this situation, $I(\mathbf{X};\mathbf{T} \mid \mathbf{S}_{oracle}) = I(\mathbf{X};\mathbf{T} \mid \mathbf{T}) = 0$, achieving the upper bound of semi-supervised learning performance.

However, in practice, without access to the ground truth $\mathbf{T}$, the gap between the fully supervised learning $\mathbf{Y}_{oracle}$ and semi-supervised learning $\mathbf{Y} \in \{\mathbf{Y}_{cr}, \mathbf{Y}_{diff}\}$ is $I(\mathbf{X};\mathbf{T} \mid \mathbf{S})$, for which we have the following inequalities:

$$\max\left(I(\mathbf{X};\mathbf{T} \mid \mathbf{S}_{cr}), I(\mathbf{X};\mathbf{T} \mid \mathbf{S}_{pr})\right) \geq \min\left(I(\mathbf{X};\mathbf{T} \mid \mathbf{S}_{cr}), I(\mathbf{X};\mathbf{T} \mid \mathbf{S}_{pr})\right) \geq I(\mathbf{X};\mathbf{T} \mid \mathbf{S}_{cr}, \mathbf{S}_{pr}). \tag{31}$$

Furthermore, based on the Equation 30, we arrive at the inequalities on the target mutual information:

$$I(\mathbf{Y}_{oracle};\mathbf{T}) \geq I(\mathbf{Y}_{diff};\mathbf{T}) \geq I(\mathbf{Y}_{cr};\mathbf{T}), \tag{32}$$

which completes the proof.

□

L DETAILED ILLUSTRATION OF DIFFMATCH FRAMEWORK

In this section, we provide detailed illustrations for a clearer understanding of the DiffMatch framework and the conditional discrete data generation pipeline using the diffusion process strategy. Figure 14 presents a comprehensive overview of the key components in DiffMatch, including the feature extractor, mask denoiser, and the supervised and unsupervised loss calculations. Figure 15 further illustrates the forward and reverse diffusion processes employed in the conditional discrete data generation pipeline for semi-supervised semantic segmentation.

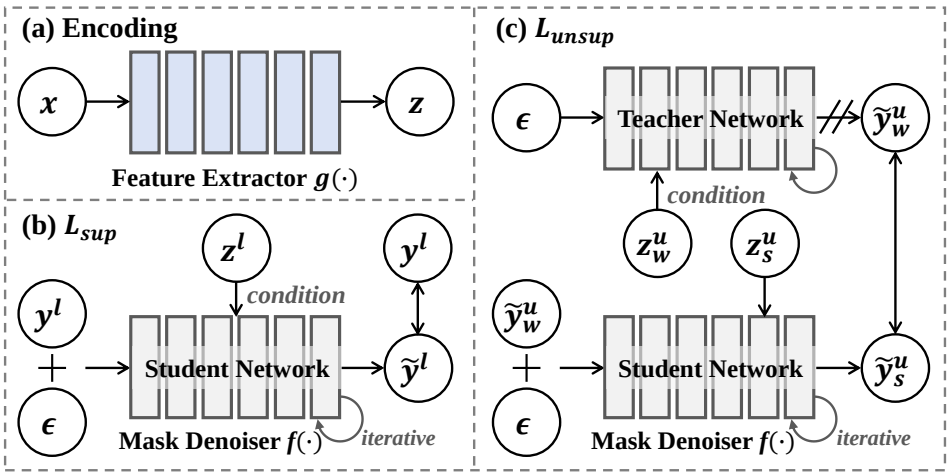


Figure 14: Illustration of the DiffMatch framework. (a) The encoding process. The feature extractor $g(\cdot)$ takes an input image $\boldsymbol{x}$ and outputs the pixel embedding $\boldsymbol{z}$. (b) Supervised loss calculation. The ground truth mask $\boldsymbol{y}^l$ is corrupted with noise $\boldsymbol{\epsilon}$ sampled from the Gaussian distribution to obtain the noisy mask $\boldsymbol{y}^l_t$. The mask denoiser $f(\cdot)$ takes $\boldsymbol{y}^l_t$ and $\boldsymbol{z}^l$ as inputs to predict the denoised mask $\tilde{\boldsymbol{y}}^l$. The supervised loss $\mathcal{L}_{sup}$ is computed between $\tilde{\boldsymbol{y}}^l$ and $\boldsymbol{y}^l$. (c) Unsupervised loss calculation. Weak and strong augmentations are applied to the unlabeled image $\mathbf{x}^u$ to obtain $\boldsymbol{x}^u_w$ and $\boldsymbol{x}^u_s$. The teacher network generates pseudo labels $\tilde{\boldsymbol{y}}^u_{0,w}$ by denoising $\boldsymbol{\epsilon}$ conditioned on $\boldsymbol{z}^u_w$. Noise is injected into $\tilde{\boldsymbol{y}}^u_{0,w}$ to obtain $\tilde{\boldsymbol{y}}^u_{t,w}$. The student network denoises $\tilde{\boldsymbol{y}}^u_{t,w}$ conditioned on $\boldsymbol{z}^u_s$ to predict $\tilde{\boldsymbol{y}}^u_s$. The unsupervised loss $\mathcal{L}_{unsup}$ is calculated between $\tilde{\boldsymbol{y}}^u_s$ and $\tilde{\boldsymbol{y}}^u_{0,w}$.

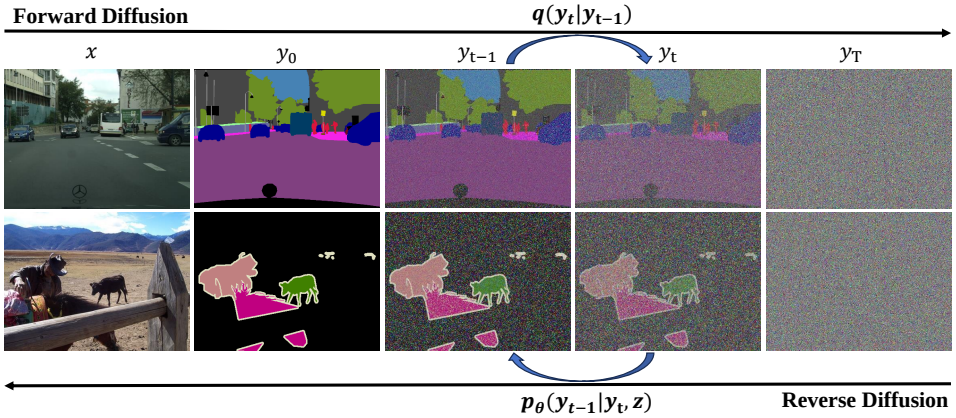


Figure 15: Illustration of the conditional discrete data generation pipeline for semi-supervised semantic segmentation using the diffusion process strategy. The forward diffusion process $q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})$ progressively corrupts the input mask $\boldsymbol{y}_0$ by adding Gaussian noise at each time step t, resulting in the noisy mask $\boldsymbol{y}_t$. The reverse diffusion process $p_\theta(\boldsymbol{y}_{t-1}|\boldsymbol{y}_t, \boldsymbol{z})$ learns to denoise the noisy mask $\boldsymbol{y}_t$ conditioned on the pixel embedding $\boldsymbol{z}$ to recover the mask $\boldsymbol{y}_{t-1}$ at previous time step. The denoising is performed iteratively, with the mask denoiser $f(\cdot)$ predicting the denoised mask at each step.