A Comprehensive Survey on Semi-Supervised Learning

1 Introduction to Semi-Supervised Learning

1.1 Introduction to Semi-Supervised Learning

Machine learning has seen tremendous success in a wide range of applications, from computer vision to natural language processing. However, the vast majority of these achievements have been made in the supervised learning setting, where a large amount of labeled data is available for training. In many real-world scenarios, obtaining labeled data can be a tedious, expensive, and time-consuming process, often requiring domain expertise or specialized knowledge. This limitation has led to the emergence of semi-supervised learning, a paradigm that aims to leverage both labeled and unlabeled data to improve the performance of machine learning models.

Semi-supervised learning (SSL) is a machine learning approach that utilizes a small amount of labeled data along with a large pool of unlabeled data to train models that can generalize better than those trained solely on labeled data [1]. The key idea behind SSL is that the underlying data distribution, which can be characterized by the unlabeled samples, can provide valuable information to guide the learning process and improve the model's performance, even when the labeled data is scarce.

Compared to supervised learning, where the model is trained solely on labeled data, and unsupervised learning, where the model is trained on unlabeled data without any label information, semi-supervised learning sits in the middle, utilizing both labeled and unlabeled data to learn more effective representations [2]. This combination of labeled and unlabeled data can lead to significant performance gains, as the unlabeled data can provide valuable insights about the underlying structure of the data, which can then be used to guide the learning process and improve the model's generalization capabilities.

The motivation behind SSL is straightforward: while labeled data is often limited and costly to obtain, unlabeled data is abundant and readily available in many domains. By exploiting the structure and patterns present in the unlabeled data, SSL algorithms can learn more robust and generalizable representations, leading to improved performance on various tasks [3]. This is particularly important in scenarios where the labeled data is limited, imbalanced, or biased, as SSL can help mitigate these issues by leveraging the additional information present in the unlabeled data.

One of the key challenges in semi-supervised learning is the effective utilization of the unlabeled data. Unlike labeled data, which provides direct supervision, unlabeled data does not have explicit labels or targets, making it more difficult to incorporate into the learning process. SSL algorithms need to make certain assumptions about the structure of the data, such as the cluster assumption (similar instances should have similar labels) or the manifold assumption (data points lie on a low-dimensional manifold), to leverage the unlabeled data effectively [4]. The success of SSL methods often depends on the validity of these assumptions and their ability to capture the underlying data distribution.

Another challenge in semi-supervised learning is the potential for the unlabeled data to introduce noise or bias into the learning process. If the unlabeled data is not representative of the true data distribution or contains out-of-distribution samples, the SSL algorithm may learn features or representations that do not generalize well to the test data [5]. Developing robust SSL methods that can handle these challenges and effectively leverage the unlabeled data is an active area of research [6].

Despite these challenges, semi-supervised learning has shown promising results in a variety of domains, including computer vision, natural language processing, and medical imaging [7]. By combining the strengths of supervised and unsupervised learning, SSL algorithms can achieve better performance with fewer labeled samples, making them a valuable tool in applications where labeled data is scarce or expensive to obtain.

1.2 Definitions and Notation

Semi-supervised learning is a machine learning paradigm that lies between supervised and unsupervised learning. In the supervised learning setting, the learner is provided with a dataset of labeled samples, where each sample is associated with a ground-truth label. The goal is to learn a function that can accurately predict the labels of new, unseen samples. In contrast, unsupervised learning aims to discover the underlying structure or patterns in the data without any label information [8].

In the semi-supervised learning setting, the learner is provided with a dataset that consists of both labeled and unlabeled samples. The labeled samples follow the supervised learning setup, while the unlabeled samples have no associated labels. The goal of semi-supervised learning is to leverage the unlabeled data, in addition to the labeled data, to learn a more accurate model compared to using only the labeled data [9].

The key parameters that characterize the semi-supervised learning setup are the number of labeled samples (n_l), the number of unlabeled samples (n_u), and the underlying data distribution $P(X, Y)$, where X and Y denote the input features and labels, respectively. The data distribution can be further decomposed into the marginal distribution of the input features, $P(X)$, and the conditional distribution of the labels given the inputs, $P(Y|X)$ [10].

In the semi-supervised learning literature, various assumptions are made about the data distribution to facilitate the effective use of unlabeled data. For instance, the cluster assumption [3] states that the decision boundaries of the classifier should lie in low-density regions of the input space. The manifold assumption [9] suggests that the data lies on a low-dimensional manifold embedded in the high-dimensional input space, and the labels can be inferred from the intrinsic geometry of this manifold.

The loss function for semi-supervised learning typically combines a supervised loss on the labeled samples and an unsupervised loss on the unlabeled samples. The supervised loss, denoted as $\mathcal{L}_s$, measures the discrepancy between the predicted labels and the ground-truth labels for the labeled samples. The unsupervised loss, denoted as $\mathcal{L}_u$, captures the underlying structure of the unlabeled data, such as the cluster or manifold assumptions. The overall semi-supervised learning objective can be written as:

$\min_{f} \mathcal{L}_s(\mathcal{L}) + \lambda \mathcal{L}_u(\mathcal{L}, \mathcal{U})$

where λ is a hyperparameter that controls the trade-off between the supervised and unsupervised losses [8].

The key challenge in semi-supervised learning is to effectively leverage the unlabeled data to improve the learning performance, especially when the number of labeled samples is limited. Various techniques have been proposed in the literature to address this challenge, such as graph-based methods [11], generative models [12], consistency regularization [13], and pseudo-labeling [14]. The survey will provide a comprehensive overview of these fundamental techniques and their recent advancements.

1.3 Assumptions and Principles of Semi-Supervised Learning

The success of semi-supervised learning (SSL) hinges on several key assumptions and principles that underlie the various techniques proposed in the literature. These assumptions and principles guide the development of SSL methods and shape their strengths, limitations, and domains of applicability.

One of the fundamental assumptions in SSL is the cluster assumption [15]. This assumption suggests that data points belonging to the same cluster or high-density region in the input space are likely to share the same label or class, and the decision boundaries of the classification task should lie in low-density regions of the data distribution. SSL methods that leverage the cluster assumption, such as graph-based approaches, aim to learn a classifier that respects the underlying cluster structure of the data, even in the presence of limited labeled samples.

Closely related to the cluster assumption is the manifold assumption [16; 17]. This assumption states that high-dimensional data often lie on or near a lower-dimensional manifold embedded in the input space, and SSL methods that build upon this assumption seek to learn a smooth classification or regression function that varies slowly along the manifold while changing rapidly across the low-density regions separating different manifolds. This principle is particularly important for high-dimensional data, where the intrinsic dimensionality is much lower than the ambient dimension.

Another important assumption in SSL is the low-density separation [18; 4]. This assumption stipulates that the decision boundaries of the classification task should lie in the low-density regions of the data distribution, where the data points are sparse. SSL methods that leverage the low-density separation, such as entropy minimization and consistency regularization, aim to push the decision boundaries away from high-density regions, thereby encouraging the classifier to make confident predictions in low-density areas.

In addition to these assumptions, semi-supervised learning with generative models [19; 20] has gained significant attention in recent years. These methods assume that the data can be generated by an underlying probabilistic model, which can be learned using both labeled and unlabeled data. The generative model component aims to capture the underlying data distribution, while the discriminative component focuses on learning the classification or regression task. The interplay between these two components, often within a unified framework, allows SSL methods to effectively leverage the information contained in the unlabeled data.

The assumptions and principles discussed above are not mutually exclusive, and many SSL methods combine various aspects to achieve better performance. The choice of which assumptions to utilize often depends on the characteristics of the data, the nature of the learning task, and the available computational resources. However, the validity and robustness of these assumptions in real-world scenarios are subject to ongoing research and debate, and violations of these assumptions can limit the effectiveness of SSL methods and lead to suboptimal performance. Consequently, recent advancements in SSL have focused on developing more flexible and robust techniques that can handle such challenges.

1.4 Evaluation Protocols and Benchmark Datasets

The evaluation of semi-supervised learning (SSL) methods is a crucial aspect of this field, as it directly impacts our understanding of the effectiveness and generalizability of these techniques. This subsection will review the commonly used evaluation protocols and benchmark datasets in the SSL literature, and discuss the importance of proper evaluation and the limitations of existing benchmarks.

One of the key challenges in evaluating SSL methods is the need for both labeled and unlabeled data. The standard evaluation protocol for SSL typically involves splitting the available data into a small labeled set and a larger unlabeled set [21]. This data split is often done randomly, which can lead to biased results as the labeled and unlabeled data may not be representative of the true data distribution [22]. To address this issue, some studies have proposed alternative evaluation protocols that take into account the distribution of the labeled and unlabeled data [23].

In addition to the data split, the choice of evaluation metrics is also crucial in assessing the performance of SSL methods. While accuracy is a commonly used metric, it may not capture the full picture, especially in scenarios where the labeled and unlabeled data have different class distributions [24]. Alternative metrics, such as balanced accuracy, have been proposed to better evaluate the fairness and robustness of SSL methods [22].

The SSL literature has relied on several benchmark datasets to evaluate the performance of these methods, including CIFAR-10, CIFAR-100, SVHN, and ImageNet [21; 25]. While these datasets have been valuable in advancing the field, they may not capture the full complexity of real-world scenarios, where the labeled and unlabeled data may come from different distributions or have different class representations [23; 26].

Moreover, the existing benchmarks often suffer from limitations in the way they handle hyperparameter tuning and the size of the labeled and unlabeled datasets [27]. For example, some studies use a labeled validation set that is much larger than the training set, which may lead to overly optimistic results [21]. To address these limitations, recent work has proposed more realistic evaluation protocols, such as using a small labeled validation set and considering the impact of domain shift between the labeled and unlabeled data [28; 29].

In summary, the evaluation of SSL methods is a complex and multifaceted problem, and the choice of evaluation protocols and benchmark datasets can have a significant impact on the reported performance of these methods. Researchers in this field must be mindful of the limitations of existing benchmarks and strive to develop more realistic and comprehensive evaluation frameworks to ensure the validity and generalizability of their findings.

2 Fundamental Techniques in Semi-Supervised Learning

2.1 Graph-based Methods

Graph-based semi-supervised learning has emerged as a powerful paradigm that can effectively leverage both labeled and unlabeled data to learn robust predictive models. These methods represent the data as a graph, where nodes correspond to data samples and edges capture the similarity or proximity between them. The key idea is to propagate the label information from the labeled nodes to the unlabeled nodes through the graph structure, thereby exploiting the intrinsic data manifold and promoting smoothness of the classifier along high-density regions [30].

One of the fundamental graph-based semi-supervised learning approaches is graph regularization, where the objective function includes a graph-based regularizer in addition to the standard supervised loss on the labeled data. The graph regularizer encourages the model's predictions to be smooth over the graph, such that nearby nodes in the graph have similar predictions. This is typically achieved by minimizing the graph Laplacian of the model's outputs, which captures the variation of the predictions across the graph [30]. The graph Laplacian can be constructed using various similarity measures, such as the Gaussian kernel or the k-nearest neighbor graph.

Another family of graph-based semi-supervised learning methods are graph embedding techniques, which aim to learn low-dimensional node representations that preserve the underlying graph structure. These methods typically start by constructing a similarity graph from the data, and then optimize an objective function that encourages nearby nodes in the graph to have similar embeddings. Once the node embeddings are learned, they can be used as features for downstream supervised tasks, such as classification or regression [11].

A prominent example of graph embedding-based semi-supervised learning is the label propagation algorithm, which iteratively updates the node embeddings by propagating label information from labeled nodes to unlabeled nodes through the graph. The label propagation process is governed by the graph structure, such that nodes that are closely connected in the graph are more likely to have similar labels [11]. The final node embeddings can then be used to train a supervised classifier on the labeled data.

One of the key advantages of graph-based semi-supervised learning methods is their ability to capture the intrinsic manifold structure of the data, which can be particularly beneficial when the labeled and unlabeled data share similar underlying distributions. However, these methods also have some limitations. First, the performance of graph-based methods can be sensitive to the choice of the graph construction and the similarity measure used [30]. Inappropriate graph construction can lead to suboptimal label propagation and poor generalization. Second, graph-based methods can be computationally expensive, especially for large-scale datasets, as they often require the computation and storage of the graph Laplacian or the pairwise similarity matrix.

To address these limitations, recent research has explored ways to improve the robustness and scalability of graph-based semi-supervised learning. For example, [30] proposed a novel regularization approach that involves a centering operation to address the high-dimensional learning inefficiency of traditional graph-based methods. [11] provided a comprehensive taxonomy of graph-based semi-supervised learning methods and highlighted promising future research directions, such as the integration of graph-based techniques with deep neural networks and the development of more efficient graph construction algorithms.

2.2 Generative Models

Generative models have played a prominent role in semi-supervised learning, leveraging both labeled and unlabeled data to learn powerful representations and generate high-quality synthetic samples. Among the popular generative approaches used in this context, variational autoencoders (VAEs) and generative adversarial networks (GANs) have been extensively explored.

VAEs are a class of generative models that learn a latent representation of the data by encoding the input into a low-dimensional latent space and then decoding the latent representation to reconstruct the original input. In the semi-supervised setting, VAEs can effectively leverage both labeled and unlabeled data to learn more informative latent representations. For example, [12] proposes a VAE-based model that can seamlessly interpolate between unsupervised, semi-supervised, and fully supervised learning by introducing a classification layer connected to the topmost encoder layer and combined with the resampled latent layer for the decoder. This approach allows the model to benefit from both the unlabeled data, which can boost unsupervised tasks, and the labeled data, which can improve classification performance.

On the other hand, GANs are a class of generative models that learn to generate realistic samples by pitting a generator network against a discriminator network in an adversarial training process. In the semi-supervised setting, GANs can be leveraged to generate high-quality synthetic samples that can augment the limited labeled data, thereby improving the performance of the classifier. [31] demonstrates the effectiveness of using GANs for semi-supervised learning, where the discriminator network is trained not only to distinguish real samples from fake samples but also to classify the labeled data, thereby learning a robust feature representation that can be used for downstream tasks.

Beyond VAEs and GANs, energy-based models (EBMs) have also been explored in the context of semi-supervised learning. EBMs define an energy function that assigns low energy to high-probability regions of the data distribution, and they can be used to learn complex, multimodal distributions without the need for explicit density estimation. In the semi-supervised setting, EBMs can be used to identify high-quality pseudo-labels for the unlabeled data by leveraging the energy scores or uncertainty estimates provided by the model. For instance, [5] shows that by inserting a small fraction of maliciously crafted unlabeled examples, the performance of semi-supervised learning models based on EBMs can be significantly degraded, highlighting the importance of developing robust semi-supervised learning techniques that can handle noisy or adversarial unlabeled data.

The trade-offs between different generative approaches in semi-supervised learning are closely tied to their underlying assumptions and modeling capabilities. VAEs, with their explicit representation of the data distribution, can provide more interpretable latent representations and are generally more stable to train. However, they may struggle to capture complex, multimodal distributions, which can be better handled by the adversarial training of GANs. EBMs, on the other hand, offer a more flexible and powerful framework for modeling complex data distributions, but they can be more challenging to train and may require careful regularization to avoid overfitting.

Overall, the use of generative models in semi-supervised learning has been a fruitful area of research, with the ability to leverage both labeled and unlabeled data to learn more robust and generalizable representations. As the field continues to evolve, we can expect to see further advancements in the integration of different generative approaches, as well as their combination with other semi-supervised learning techniques, to address the growing challenges in real-world applications.

2.3 Consistency Regularization

Consistency regularization is a fundamental technique in semi-supervised learning that aims to enforce the model's predictions to be consistent or invariant to small perturbations of the input data. The key idea behind consistency regularization is the cluster assumption, which posits that the decision boundaries of the classifier should lie in low-density regions of the data distribution [3]. By encouraging the model to make similar predictions for input samples that are close in the input space, consistency regularization can effectively leverage the structure of the unlabeled data to improve the model's performance.

One of the popular consistency regularization methods is virtual adversarial training (VAT) [32]. VAT perturbs the input with a small adversarial perturbation and then minimizes the KL divergence between the model's predictions on the original and perturbed inputs, effectively enforcing the consistency of the model's predictions. By generating the adversarial perturbation in an unsupervised manner, VAT can leverage the structure of the unlabeled data to improve the model's robustness and generalization.

Building upon the success of VAT, Mixup [33] has emerged as another influential consistency regularization method. Mixup generates new training examples by linearly interpolating pairs of input samples and their corresponding labels. This encourages the model to make smooth predictions between data points, which can be particularly effective in semi-supervised learning where the unlabeled data can provide valuable information about the data distribution.

More recently, FixMatch [34] has been proposed as a state-of-the-art consistency regularization method for semi-supervised learning. FixMatch combines two key ideas: (1) weak augmentation, which applies a simple data augmentation technique such as random cropping and flipping to the unlabeled data, and (2) strong augmentation, which applies a more aggressive data augmentation technique such as RandAugment or CutOut to the same unlabeled data. FixMatch then enforces consistency between the model's predictions on the weakly augmented and strongly augmented unlabeled samples, effectively leveraging the structure of the unlabeled data to improve the model's performance.

The effectiveness of consistency regularization techniques in semi-supervised learning can be attributed to their ability to exploit the underlying data manifold and enforce smooth decision boundaries. By encouraging the model to make similar predictions for nearby samples, consistency regularization can effectively propagate the information from the labeled data to the unlabeled data, leading to significant performance improvements, especially in the low-label regime. Moreover, consistency regularization methods have been shown to be particularly effective when combined with other semi-supervised learning techniques, such as pseudolabeling and graph-based methods [35], further enhancing the model's ability to leverage the unlabeled data and improve its generalization performance.

Overall, consistency regularization is a powerful and widely adopted technique in the semi-supervised learning literature. Its ability to effectively exploit the structure of the data distribution and enforce smooth decision boundaries has made it a key component in many state-of-the-art semi-supervised learning algorithms. As the field of semi-supervised learning continues to evolve, we can expect to see further advancements and refinements in consistency regularization techniques, leading to even more effective and robust models for leveraging unlabeled data.

2.4 Pseudo-labeling and Self-training

Pseudo-labeling and self-training are powerful semi-supervised learning (SSL) techniques that leverage unlabeled data to improve model performance. These approaches generate predicted labels, or pseudo-labels, for the unlabeled data and then use them to iteratively refine the model.

The pseudo-labeling approach involves training an initial model on the limited labeled data and then using that model to predict labels for the unlabeled data. The high-confidence predictions are then used as pseudo-labels to augment the training data and retrain the model, aiming to gradually improve the model's performance [36]. Techniques such as confidence-based thresholding have been proposed to improve the quality of pseudo-labels by only selecting high-confidence predictions, mitigating the risk of introducing noisy or incorrect labels [37]. Moreover, mutual learning approaches, where multiple models are trained simultaneously and learn from each other's predictions, have been shown to enhance the robustness of pseudo-labeling [38].

In addition, the use of contrastive learning has emerged as a way to further improve the quality of pseudo-labels. Contrastive pseudo-labeling techniques leverage the concept of contrastive learning to identify high-quality pseudo-labels by enforcing consistency between the predictions of differently augmented versions of the same unlabeled sample [39]. This approach helps to capture the intrinsic similarities and differences within the unlabeled data, leading to more reliable pseudo-labels.

Self-training, a closely related paradigm, takes the pseudo-labeling concept a step further by incorporating the predictions of the model itself into the training process. In self-training, the model is first trained on the limited labeled data, and then used to generate predictions for the unlabeled data. The most confident predictions are then added to the training set, and the model is retrained on the augmented dataset [40]. This iterative process aims to gradually improve the model's performance by incorporating the informative, yet unlabeled, data samples.

The success of pseudo-labeling and self-training approaches relies heavily on the quality of the generated pseudo-labels. Techniques that address the challenges of noisy pseudo-labels, class imbalance, and out-of-distribution samples have emerged as important research directions. For example, [41] proposes a method that identifies the best pseudo-label allocation via optimal transport to only samples with high confidence scores, mitigating the impact of unreliable pseudo-labels.

Furthermore, the integration of pseudo-labeling and self-training with other semi-supervised learning techniques, such as consistency regularization and generative models, has shown promising results. By leveraging the complementary strengths of these approaches, researchers have developed more robust and effective semi-supervised learning frameworks [42].

Overall, pseudo-labeling and self-training are powerful techniques that have been extensively explored in the semi-supervised learning literature. Ongoing research focuses on improving the quality and reliability of pseudo-labels, as well as developing more effective ways to incorporate the unlabeled data into the training process. As the field of semi-supervised learning continues to evolve, the advancements in pseudo-labeling and self-training are likely to play a crucial role in bridging the gap between limited labeled data and abundant unlabeled data in real-world applications.

3 Advances in Semi-Supervised Learning

3.1 Leveraging Self-Supervised Learning

The recent advancements in self-supervised learning have demonstrated the potential to enhance semi-supervised learning. Self-supervised learning aims to learn useful representations from the structure and patterns present in unlabeled data, without requiring manual annotation. These self-supervised representations can then be leveraged to improve the performance of semi-supervised learning, especially in scenarios where labeled data is scarce.

One of the key techniques that has shown promising results in this direction is the use of contrastive self-supervised learning [43; 44]. Contrastive learning aims to learn representations by enforcing similar representations for augmented versions of the same input, while pushing apart representations of different inputs. This principle can be directly applied to semi-supervised learning, where the labeled data can be used to guide the contrastive learning objective on the unlabeled data.

A popular approach in this direction is to first pre-train a network using contrastive self-supervised learning on the unlabeled data, and then fine-tune the network using the limited labeled data for the target task [45; 46]. The self-supervised pre-training helps learn robust and generalizable representations, which can then be effectively fine-tuned using the labeled data. This two-stage approach has demonstrated significant improvements over training the model solely on the limited labeled data.

Another line of work has focused on integrating the self-supervised and semi-supervised objectives in a more seamless manner. For example, [47] proposes a unified framework called Self-Tuning that combines the exploration of labeled and unlabeled data with the transfer of a pre-trained model. The key idea is to leverage the pre-trained model to provide an implicit regularization, while simultaneously using the self-supervised objective to extract useful information from the unlabeled data. This helps mitigate the dilemma of needing a decent pre-trained model for transfer learning, or having a large pool of unlabeled data for effective semi-supervised learning.

More recently, [48] has explored the use of generative foundation models, trained on large-scale datasets, to generate synthetic unlabeled samples for semi-supervised learning. The key insight is that the representations learned by these generative models can provide a useful substitute for real unlabeled data, especially in scenarios where access to unlabeled data is limited due to practical constraints. The method formulates an alternating optimization problem to jointly meta-learn the generative model and train the semi-supervised classifier, demonstrating the potential of leveraging large-scale self-supervised models for data-efficient semi-supervised learning.

Overall, the integration of self-supervised learning has emerged as a promising direction to enhance the effectiveness of semi-supervised learning, particularly in scenarios where labeled data is scarce. The ability of self-supervised models to learn robust and generalizable representations from unlabeled data can be effectively leveraged to boost the performance of semi-supervised classifiers trained on limited labeled data. As the field of self-supervised learning continues to evolve, we can expect to see further advancements in the synergistic integration of self-supervised and semi-supervised learning techniques.

3.2 Advancements in Meta-Learning for Semi-Supervised Learning

Meta-learning, also known as learning to learn, has emerged as a promising paradigm in the field of semi-supervised learning. Unlike traditional machine learning approaches that optimize a model for a specific task using a fixed dataset, meta-learning algorithms aim to learn how to efficiently learn new tasks or adapt to new data distributions with limited labeled samples [49]. This ability to learn learning strategies can be particularly valuable in the context of semi-supervised learning, where the goal is to leverage both labeled and unlabeled data to improve the model's performance.

One prominent approach in this direction is model-agnostic meta-learning (MAML), which learns an initialization of the model parameters that can be quickly adapted to new tasks using a small number of samples [49]. In the semi-supervised setting, MAML can be extended to learn an initialization that can be efficiently fine-tuned using both labeled and unlabeled data. For example, the Meta-Semi-Supervised Learning (Meta-SSL) framework [50] proposes a meta-learning objective that encourages the model to learn representations that can be easily adapted to new tasks using a combination of labeled and unlabeled data.

Another line of work has focused on meta-learning strategies that explicitly optimize the model's ability to leverage unlabeled data. The Adversarial Meta-Learning (AML) algorithm [51] trains a meta-learner to generate task-specific data augmentation policies that can effectively exploit unlabeled data for semi-supervised learning. The key idea is to learn an augmentation policy that, when applied to the labeled and unlabeled data, can improve the model's performance on a held-out task. By optimizing the meta-learner to generate effective augmentation policies, AML can significantly boost the performance of semi-supervised learning models, especially when the labeled data is scarce.

Similarly, the Variational Information Distillation (VID) framework [52] employs meta-learning to learn a shared representation that can effectively transfer knowledge from labeled to unlabeled data. VID trains a meta-learner to distill the information from a teacher model, which is trained on the labeled data, into a student model that can leverage both labeled and unlabeled data. By optimizing the meta-learner to learn a representation that facilitates this knowledge transfer, VID can achieve state-of-the-art results in semi-supervised learning tasks.

Beyond MAML-based approaches, researchers have also explored meta-learning strategies that learn to optimize the semi-supervised learning process itself. The Data Efficient Meta-Learning (DEML) algorithm [53] learns an optimization procedure that can efficiently fine-tune a model using a small number of labeled samples and a large amount of unlabeled data. DEML trains a meta-learner to generate the updates for the model parameters, allowing it to quickly adapt to new tasks with limited labeled data while leveraging the unlabeled samples.

Overall, the advancements in meta-learning for semi-supervised learning have demonstrated the potential of learning-to-learn approaches to address the challenges of limited labeled data. By optimizing the model's initialization, data augmentation policies, or the optimization process itself, meta-learning techniques can enable semi-supervised learning models to effectively leverage unlabeled data and achieve competitive performance with reduced annotation efforts. As the field of semi-supervised learning continues to evolve, the integration of meta-learning with other advanced techniques, such as energy-based models and self-supervised learning, is likely to yield further advancements in this important area of machine learning.

3.3 Energy-based Models for Semi-Supervised Learning

One promising avenue for advancing semi-supervised learning is the use of energy-based models (EBMs), which have shown great potential in effectively incorporating both labeled and unlabeled data. EBMs define an energy function that assigns a scalar value to each input, with lower energy values corresponding to more plausible or likely inputs. This unique property of EBMs can be leveraged in the semi-supervised setting to identify high-quality pseudo-labels from the unlabeled data while rejecting low-confidence or out-of-distribution samples.

A key advantage of EBMs is their ability to capture complex data distributions without making restrictive assumptions, unlike generative models that often struggle to model the true data distribution [19]. This flexibility allows EBMs to be trained using a variety of techniques, including maximum likelihood estimation, contrastive divergence, and adversarial training, making them a versatile framework for semi-supervised learning.

One prominent example of using EBMs for semi-supervised learning is the Confident Sinkhorn Allocation (CSA) approach [41]. CSA leverages the uncertainty estimates provided by an EBM to selectively identify high-confidence pseudo-labels from the unlabeled data, and then uses an optimal transport-based allocation to propagate these pseudo-labels to the rest of the unlabeled samples. By carefully selecting the most reliable pseudo-labels, CSA can outperform traditional pseudo-labeling methods, which are often susceptible to error propagation.

Similarly, the Semi-supervised Contrastive Outlier removal for Pseudo Expectation Maximization (SCOPE) method [54] combines the strengths of EBMs and contrastive learning to address the issue of confounding errors in pseudo-labeling. SCOPE uses an energy-based component to estimate the uncertainty of unlabeled samples and suppress the influence of low-confidence or out-of-distribution samples, while the contrastive learning component helps to learn robust representations from both labeled and unlabeled data.

Moreover, the use of EBMs in semi-supervised learning can be further enhanced by leveraging self-supervised learning techniques. The GEDI framework [55] proposes a unified approach that integrates self-supervised learning objectives with an energy-based generative model, enabling the joint training of discriminative and generative components. This approach allows the model to learn powerful representations from unlabeled data while also benefiting from the structure and uncertainty estimates provided by the energy-based component.

Another line of research explores the integration of EBMs with other semi-supervised learning paradigms, such as meta-learning. The work on Meta-Pseudo-Labels [56] combines an EBM-based uncertainty estimation with a meta-learning framework to learn an effective pseudo-labeling strategy, which can then be applied to a wide range of semi-supervised learning tasks.

The use of EBMs for semi-supervised learning is a promising direction that has shown impressive results, particularly in addressing the challenges of confounding errors and out-of-distribution samples. By leveraging the expressive power and uncertainty estimates of EBMs, researchers have been able to develop more robust and reliable semi-supervised learning methods that can effectively utilize both labeled and unlabeled data. As the field of semi-supervised learning continues to evolve, the integration of EBMs with other advanced techniques, such as self-supervised learning and meta-learning, is likely to yield further advancements in this important area of machine learning.

3.4 Adversarial Training for Semi-Supervised Learning

The integration of adversarial training techniques with semi-supervised learning has emerged as a promising approach to improve the robustness and generalization of models when using both labeled and unlabeled data. Adversarial training, which aims to enhance the model's resilience against adversarial perturbations, can be particularly beneficial in the semi-supervised setting where the unlabeled data may contain a significant amount of noise or out-of-distribution samples.

One of the key advantages of incorporating adversarial training into semi-supervised learning is its ability to learn more robust and transferable representations from the unlabeled data. [21] has shown that the performance of semi-supervised learning methods can degrade substantially when the unlabeled dataset contains out-of-class examples. Adversarial training can help mitigate this issue by encouraging the model to learn more generalizable features that are invariant to small perturbations, which are often indicative of out-of-distribution samples.

This integration of adversarial training and semi-supervised learning has been explored in several recent studies, often demonstrating significant improvements over standalone semi-supervised approaches. [29] proposes a method called RealMix that leverages adversarial training to enhance the model's ability to handle distribution shifts between the labeled and unlabeled data. By generating adversarial examples and incorporating them into the training process, RealMix is able to learn more robust representations that are less sensitive to the mismatch between the labeled and unlabeled data distributions.

Another line of research has focused on using adversarial training to improve the quality of pseudo-labels generated for the unlabeled data, which is a crucial component in many semi-supervised learning algorithms. [5] has shown that adversaries can easily manipulate the unlabeled dataset to mislead semi-supervised models by inserting maliciously-crafted unlabeled examples. To address this vulnerability, [57] introduces an "Unbiased Teacher" approach that jointly trains a student and a gradually progressing teacher model in a mutually-beneficial manner, using adversarial training to ensure the reliability of the pseudo-labels.

Beyond just improving the quality of pseudo-labels, adversarial training can also be used to enhance the consistency of the model's predictions across different perturbations of the input data, a key principle underlying many successful semi-supervised learning methods. [28] proposes a technique called "Split Batch Normalization" that uses separate batch normalization statistics for labeled and unlabeled data to improve the consistency of the model's predictions, particularly in the presence of domain shifts between the labeled and unlabeled data.

The integration of adversarial training with semi-supervised learning has also shown promising results in more complex domains, such as medical imaging. [58] leverages adversarial training to estimate the model's uncertainty, which is then used to selectively utilize the unlabeled data and improve the consistency of the model's predictions in a semi-supervised medical image segmentation task.

Overall, the combination of adversarial training and semi-supervised learning has demonstrated its ability to enhance the robustness, generalization, and reliability of models trained on limited labeled data, while effectively leveraging the abundant unlabeled data. As the field of semi-supervised learning continues to evolve, the incorporation of adversarial training techniques is likely to play an increasingly important role in addressing the challenges posed by real-world, noisy, and out-of-distribution data.

4 Applications and Benchmark Datasets

4.1 Computer Vision

Semi-supervised learning has demonstrated significant potential in various computer vision tasks, leveraging both labeled and unlabeled data to improve model performance. In the domain of image classification, a prominent example is the application of semi-supervised learning on the CIFAR benchmark dataset [2]. Several state-of-the-art semi-supervised methods, such as MixMatch [21], ReMixMatch [59], and FixMatch [34], have shown remarkable results on the CIFAR-10 and CIFAR-100 datasets, achieving accuracies of up to 95.10% and 83.91%, respectively, with only 4,000 labeled samples [60].

The success of semi-supervised learning in computer vision can be attributed to its ability to effectively leverage the rich information contained in unlabeled data, which is often abundant in real-world scenarios. By exploiting techniques such as consistency regularization, pseudo-labeling, and contrastive learning, semi-supervised methods can learn more robust and generalizable representations, outperforming their fully-supervised counterparts, especially when labeled data is scarce [21].

In the domain of semantic segmentation, semi-supervised learning has also proven to be effective. The GuidedMix-Net [61] framework leverages labeled information to guide the learning of unlabeled instances, achieving significant improvements in mIoU on the PASCAL VOC 2012 and Cityscapes datasets compared to previous semi-supervised approaches. Moreover, the DiverseNet [62] method explores multi-head and multi-model semi-supervised learning algorithms to enhance precision and diversity during the training process, leading to state-of-the-art performance on various remote sensing imagery datasets.

Semi-supervised learning has also been applied to object detection tasks, where labeled data is particularly scarce. The Unbiased Teacher [57] approach addresses the pseudo-labeling bias issue in semi-supervised object detection by jointly training a student and a gradually progressing teacher in a mutually-beneficial manner. This method has achieved significant improvements over state-of-the-art methods, with up to 6.8 absolute mAP improvements on the MS-COCO dataset when using only 1% of labeled data.

Beyond the standard benchmark datasets, semi-supervised learning has also been explored in various real-world computer vision applications. In the medical imaging domain, the Uncertainty-Aware Deep Co-training for Semi-supervised Medical Image Segmentation [58] method leverages Monte Carlo Sampling to estimate uncertainty maps and guide the training process, leading to substantial performance gains on challenging medical datasets.

However, the performance of semi-supervised learning in computer vision is not without its challenges. Recent studies have highlighted the potential for semi-supervised methods to exhibit biases, where the benefits of unlabeled data may disproportionately favor certain subpopulations over others [24]. Addressing these fairness concerns and ensuring the equitable application of semi-supervised learning in computer vision remains an important avenue for future research.

4.2 Natural Language Processing

The application of semi-supervised learning (SSL) in natural language processing (NLP) has gained significant attention in recent years, particularly with the emergence of large language models (LLMs) [63; 64]. These powerful models, trained on vast amounts of unlabeled text data, have demonstrated remarkable capabilities in various NLP tasks, including text classification, machine translation, and language modeling.

One of the key advancements in SSL for NLP is the use of unsupervised pretraining, where LLMs are first trained on large corpora of unlabeled data to learn rich, transferable representations [63; 64]. These pretrained models can then be fine-tuned on smaller, labeled datasets for specific tasks, leveraging the knowledge acquired during pretraining. This approach has been particularly effective in scenarios where labeled data is scarce, as the pretraining phase allows the model to learn general linguistic patterns and semantics from the unlabeled data.

Another important technique in SSL for NLP is self-training, where the model is iteratively trained on its own predictions on unlabeled data, gradually improving its performance [38]. This approach can be effective in tasks where the model's predictions on unlabeled data can be used as reliable pseudo-labels, such as text classification. By incorporating these pseudo-labels into the training process, the model can learn from the unlabeled data and enhance its performance on the labeled data.

Consistency regularization, a popular SSL technique, has also shown promising results in NLP applications [13]. This approach enforces the model's predictions to be invariant to small perturbations of the input text, encouraging the model to learn robust and generalizable representations. Techniques like virtual adversarial training and FixMatch have been successfully applied to NLP tasks, demonstrating the effectiveness of consistency regularization in leveraging unlabeled data.

Furthermore, the integration of SSL with other machine learning paradigms, such as few-shot learning and continual learning, has emerged as a promising direction in NLP [65]. By leveraging the synergies between these techniques, researchers aim to develop more robust and versatile models that can adapt to data-scarce scenarios and evolving data distributions.

One notable example is the application of semi-supervised meta-learning, where the model learns to effectively utilize unlabeled data by optimizing its initialization or learning strategy in a meta-learning framework [65]. This approach can be particularly beneficial in scenarios where the target task involves a different distribution or attribute space compared to the training data, as the meta-learning process can help the model generalize better to such heterogeneous settings.

Despite the significant progress in SSL for NLP, there are still open challenges and research opportunities in this domain. Addressing the robustness of SSL methods to noisy or out-of-distribution unlabeled data, developing scalable and efficient algorithms for large-scale NLP tasks, and further integrating SSL with other machine learning paradigms are some of the key areas that researchers are actively exploring [66].

4.3 Medical Imaging

The application of semi-supervised learning in medical imaging has shown significant promise, particularly in tasks such as image segmentation, disease classification, and anomaly detection. The medical domain presents unique challenges and opportunities for the application of semi-supervised techniques, as the availability of labeled data is often scarce and the consequences of incorrect predictions can be severe.

One of the key areas where semi-supervised learning has been applied in medical imaging is image segmentation. Accurate segmentation of anatomical structures or pathological regions is crucial for many clinical applications, such as disease diagnosis, treatment planning, and monitoring. However, obtaining pixel-wise annotations for medical images can be an extremely labor-intensive and time-consuming task, requiring expert radiologists or clinicians. Semi-supervised methods have been explored to leverage the vast amount of unlabeled medical images to improve the performance of segmentation models, even with limited labeled data. [58]

Another crucial application of semi-supervised learning in medical imaging is disease classification. Identifying and classifying diseases or anomalies from medical images, such as cancer detection in mammograms or lung disease diagnosis from chest X-rays, can have a significant impact on patient outcomes. However, the limited availability of labeled medical data poses a significant challenge for supervised learning approaches. Semi-supervised methods have shown the potential to leverage the abundant unlabeled data to improve the performance of disease classification models, even when the labeled dataset is small. [67]

The medical domain also presents unique challenges for semi-supervised learning, such as the presence of rare diseases, class imbalance, and the need for robust and reliable predictions. [59] addressed the issue of unreliable classifiers in semi-supervised learning by developing an approach that can provide reliable uncertainty quantification, even when the labels are missing at random.

In summary, the application of semi-supervised learning in medical imaging has demonstrated the potential to leverage the abundance of unlabeled data to enhance the performance of critical tasks like image segmentation and disease classification, particularly in scenarios with limited labeled data. As the field continues to evolve, the integration of semi-supervised learning with other emerging techniques, such as self-supervised learning and meta-learning, may further enhance the capabilities of medical imaging systems and contribute to advancements in clinical diagnosis and treatment.

4.4 Speech Recognition

The application of semi-supervised learning in speech recognition has gained significant attention in recent years, particularly in the domains of automatic speech recognition (ASR) and speaker diarization. Semi-supervised techniques have shown the potential to leverage the abundance of unlabeled speech data to improve the performance of speech recognition models, especially when labeled data is limited.

One of the key benefits of using semi-supervised learning in speech recognition is the ability to integrate it with acoustic models, which form the core of ASR systems. Acoustic models are responsible for mapping the input audio features to the underlying phonemes or words, and their performance is heavily dependent on the availability of high-quality labeled data for training. [37] has highlighted the effectiveness of using semi-supervised techniques like task-adaptive pre-training (TAPT) to enhance the acoustic models, even when the labeled data is scarce.

Similarly, in the context of speaker diarization, which aims to identify and segment different speakers within an audio recording, semi-supervised learning can play a crucial role. Speaker diarization is a challenging task, as it requires accurately detecting speaker changes and clustering speech segments belonging to the same speaker. [57] has demonstrated that semi-supervised approaches can effectively leverage unlabeled speech data to improve the performance of speaker diarization models, particularly when the labeled data is limited.

To evaluate the performance of semi-supervised speech recognition models, researchers often use benchmark datasets such as LibriSpeech, a large-scale dataset of read English speech derived from audiobooks, and the Wall Street Journal (WSJ) dataset, which contains read speech from news articles. [68] and [27] have explored the use of semi-supervised techniques on these datasets, showcasing the potential of leveraging unlabeled data to enhance the accuracy of ASR models.

While the application of semi-supervised learning in speech recognition has shown promising results, the field also faces unique challenges. [5] has highlighted the vulnerability of semi-supervised speech recognition models to data poisoning attacks, where maliciously crafted unlabeled examples can manipulate the model's performance. Addressing these challenges and developing robust semi-supervised techniques for speech recognition is an active area of research.

Furthermore, the integration of semi-supervised learning with other machine learning paradigms, such as few-shot learning and continual learning, can further enhance the performance and applicability of speech recognition systems. [69] discusses the potential of combining semi-supervised learning with these complementary approaches to address the challenges faced in real-world speech recognition scenarios, such as adapting to new speakers or domains with limited labeled data.

Overall, the application of semi-supervised learning in speech recognition has shown promising results, with the ability to leverage unlabeled speech data to improve the performance of acoustic models and speaker diarization systems. However, the field also faces unique challenges, such as data poisoning attacks and the need for further integration with other machine learning techniques. As the research in this area continues to evolve, we can expect to see more robust and versatile semi-supervised speech recognition models that can adapt to the diverse and dynamic requirements of real-world applications.

5 Theoretical Insights and Analysis

5.1 Assumptions and Theoretical Foundations

The theoretical analysis of semi-supervised learning typically relies on several key assumptions about the underlying data distribution and the relationship between labeled and unlabeled data. These assumptions serve as the foundation for deriving convergence guarantees, generalization bounds, and understanding the potential benefits of leveraging unlabeled data.

One of the most commonly invoked assumptions in semi-supervised learning is the manifold assumption [70], which posits that the high-dimensional data lies on or near a low-dimensional manifold. This assumption suggests that nearby points in the input space are likely to share the same label, and the decision boundary should lie in a low-density region of the manifold. The manifold assumption is closely related to the cluster assumption [71], which states that data points belonging to the same class are more likely to form compact clusters in the feature space, and the decision boundary should pass through low-density regions between these clusters.

Building upon these core assumptions, the theoretical analysis of semi-supervised learning often follows a general framework that aims to characterize the convergence and generalization properties of semi-supervised algorithms. This framework typically involves bounding the excess risk or the generalization error of the semi-supervised learner in terms of the number of labeled and unlabeled samples, the complexity of the hypothesis class, and the degree of satisfying the underlying assumptions [3].

Another important assumption is the smoothness assumption [72], which stipulates that the underlying function to be learned should vary smoothly along the manifold. This implies that small perturbations to the input data should not significantly change the model's predictions, and nearby points are likely to have similar labels. The smoothness assumption serves as the theoretical basis for many semi-supervised learning techniques, such as graph-based methods and consistency regularization.

For example, one line of research has derived generalization bounds for semi-supervised learning methods, showing that under certain conditions, the sample complexity required to achieve a desired level of performance can be significantly lower than the sample complexity of supervised learning [73]. These bounds often depend on the degree of satisfying the manifold assumption, the cluster assumption, or the smoothness assumption, as well as the distribution mismatch between the labeled and unlabeled data [4].

Additionally, the theoretical analysis has also explored the convergence properties of semi-supervised learning algorithms, such as the convergence rates of graph-based methods and generative models [1]. These analyses provide insights into the conditions under which semi-supervised learning can outperform supervised learning and the limitations of various semi-supervised techniques.

Overall, the theoretical foundations of semi-supervised learning provide a rigorous framework for understanding the capabilities and limitations of these methods, and guide the development of more effective semi-supervised learning algorithms. By carefully examining the underlying assumptions and their implications, researchers can gain valuable insights into the practical applicability of semi-supervised learning and identify promising directions for further advancements in this field.

5.2 Generalization Bounds and Sample Complexity

The theoretical analysis of semi-supervised learning typically relies on several key assumptions about the underlying data distribution and the relationship between labeled and unlabeled data. These assumptions serve as the foundation for deriving convergence guarantees, generalization bounds, and understanding the potential benefits of leveraging unlabeled data.

One of the most commonly invoked assumptions in semi-supervised learning is the manifold assumption [70], which posits that the high-dimensional data lies on or near a low-dimensional manifold. This assumption suggests that nearby points in the input space are likely to share the same label, and the decision boundary should lie in a low-density region of the manifold. The manifold assumption is closely related to the cluster assumption [71], which states that data points belonging to the same class are more likely to form compact clusters in the feature space, and the decision boundary should pass through low-density regions between these clusters.

The theoretical analysis of semi-supervised learning often follows a general framework that aims to characterize the convergence and generalization properties of semi-supervised algorithms. This framework typically involves bounding the excess risk or the generalization error of the semi-supervised learner in terms of the number of labeled and unlabeled samples, the complexity of the hypothesis class, and the degree of satisfying the underlying assumptions [3].

Another important assumption is the smoothness assumption [72], which stipulates that the underlying function to be learned should vary smoothly along the manifold. This implies that small perturbations to the input data should not significantly change the model's predictions, and nearby points are likely to have similar labels. The smoothness assumption serves as the theoretical basis for many semi-supervised learning techniques, such as graph-based methods and consistency regularization.

The theoretical analysis has also explored the convergence properties of semi-supervised learning algorithms, such as the convergence rates of graph-based methods and generative models [1]. These analyses provide insights into the conditions under which semi-supervised learning can outperform supervised learning and the limitations of various semi-supervised techniques.

Overall, the theoretical foundations of semi-supervised learning provide a rigorous framework for understanding the capabilities and limitations of these methods, and guide the development of more effective semi-supervised learning algorithms. By carefully examining the underlying assumptions and their implications, researchers can gain valuable insights into the practical applicability of semi-supervised learning and identify promising directions for further advancements in this field.

5.3 Analysis of Specific Semi-Supervised Learning Techniques

This subsection will provide a detailed theoretical analysis of the key semi-supervised learning techniques, such as graph-based methods, generative models, consistency regularization, pseudo-labeling, and self-training.

Graph-based semi-supervised learning methods rely on the assumption that data points lying close to each other on the underlying data manifold should have similar predictions. This is formalized through the use of a graph Laplacian regularizer, which encourages the learned classifier to be smooth with respect to the graph structure. The theoretical analysis of these methods typically involves investigating the properties of the graph Laplacian operator, such as its spectral decomposition and the relationship between the graph structure and the data distribution. For example, [17] provides a formal justification for graph-based semi-supervised learning by showing that under certain assumptions, the bandlimited interpolation of graph signals is closely related to a constrained low-density separation problem.

Generative model-based semi-supervised learning approaches rely on the assumption that the data can be well-modeled by a generative process, which can be leveraged to learn from both labeled and unlabeled data. The theoretical analysis of these methods often focuses on the properties of the generative model, such as the expressiveness of the model family, the identifiability of the model parameters, and the convergence of the learning algorithms. For instance, [74] provides a theoretical analysis of the semi-supervised learning framework that combines manifold regularization with data representation methods, such as non-negative matrix factorization and sparse coding.

Consistency regularization techniques aim to enforce the model's predictions to be invariant to small perturbations of the input data. The theoretical analysis of these methods often focuses on the properties of the consistency loss function, the assumptions required for the success of the approach, and the convergence guarantees. For example, [75] provides a theoretical and empirical analysis of the limitations of using entropy-based uncertainty for multi-class semi-supervised segmentation tasks.

Pseudo-labeling and self-training approaches iteratively generate predicted labels for unlabeled data and use them to refine the model. The theoretical analysis of these methods typically investigates the assumptions required for the success of the approach, the reliability of the pseudo-labels, and the convergence properties of the iterative training process. For instance, [59] provides a theoretical framework for semi-supervised learning with missing labels, which can be applied to pseudo-labeling and self-training methods.

Overall, the theoretical analysis of the key semi-supervised learning techniques often involves investigating the underlying assumptions, convergence properties, and generalization bounds of the respective approaches. By understanding the theoretical foundations of these methods, researchers can gain insights into their strengths, limitations, and the potential for further improvements.

5.4 Role of Data Distribution and Model Complexity

The performance of semi-supervised learning algorithms is heavily dependent on the underlying data distribution and the complexity of the prediction model. The characteristics of the data, such as the degree of cluster separation, the manifold structure, and the presence of outliers, can significantly affect the effectiveness of different semi-supervised approaches.

Several studies have highlighted the importance of the data distribution in the success of semi-supervised learning. For instance, [73] shows that semi-supervised learning can lead to significant improvements in the learning rate compared to supervised learning, but only under specific conditions on the data distribution. The authors demonstrate that if the unlabeled data is distributed in a way that satisfies certain assumptions, such as having well-separated clusters, semi-supervised learning can achieve faster convergence rates. Conversely, if the data distribution does not satisfy these assumptions, the performance of semi-supervised learning may be worse than that of supervised learning.

Similarly, [69] investigates the role of the data distribution in the success of semi-supervised learning. The authors show that semi-supervised learning can only work effectively when the problem is "well-conditioned," meaning that the unlabeled data provides sufficient information to infer the class labels. They argue that this is a reasonable but strong assumption, and that semi-supervised learning may not be universally applicable to all learning problems.

The manifold structure of the data is another crucial factor that can impact the performance of semi-supervised learning. [3] discusses the importance of the manifold assumption, which states that the data lies on a low-dimensional manifold embedded in the high-dimensional input space. When this assumption holds, semi-supervised learning methods that leverage the manifold structure, such as graph-based approaches, can be effectively applied to exploit the unlabeled data. However, if the data does not exhibit a clear manifold structure, these methods may not provide significant improvements over supervised learning.

Furthermore, the presence of outliers or noise in the data can also hinder the performance of semi-supervised learning. [5] demonstrates that adversarial attacks on the unlabeled dataset can severely degrade the performance of semi-supervised learning models, highlighting the importance of data quality and robustness in these settings.

In addition to the data distribution, the complexity of the prediction model also plays a crucial role in the success of semi-supervised learning. [37] shows that the effectiveness of data re-sampling techniques, which are commonly used in semi-supervised learning to handle class imbalance, depends on the stage of the training process. Specifically, the authors find that re-sampling can be beneficial for training the classifier, but it can actually harm the training of the feature extractor. This suggests that the interplay between the model complexity and the amount of labeled and unlabeled data is an important consideration in semi-supervised learning.

Overall, the performance of semi-supervised learning is heavily influenced by the characteristics of the data distribution, such as the degree of cluster separation, the manifold structure, and the presence of outliers. Additionally, the complexity of the prediction model and the amount of labeled and unlabeled data required for effective learning are also important factors to consider. Understanding these dependencies is crucial for the successful application of semi-supervised learning in real-world scenarios.

6 Trends and Future Directions

6.1 Integration of Semi-Supervised Learning with Other Machine Learning Paradigms

The synergistic integration of semi-supervised learning with other machine learning paradigms holds immense potential in addressing the challenges posed by data-scarce scenarios and evolving data distributions. By leveraging the complementary strengths of various learning approaches, researchers can develop more robust, versatile, and data-efficient models that can thrive in challenging real-world settings.

One promising direction is the intersection of semi-supervised learning and few-shot learning [76]. Few-shot learning aims to learn effective models from a limited number of labeled examples, making it a natural complement to semi-supervised learning. The key idea is to utilize the unlabeled data in the semi-supervised setting to learn rich and transferable representations, which can then be effectively fine-tuned on the few labeled examples in a few-shot learning scenario, enabling quick adaptation to new tasks or domains.

Another exciting direction is the integration of semi-supervised learning with self-supervised learning. Recent advancements in self-supervised techniques, such as contrastive learning and masked language modeling, have demonstrated the ability to learn powerful representations from unlabeled data [25]. By seamlessly combining self-supervised pre-training with semi-supervised fine-tuning, researchers can leverage the complementary strengths of these approaches to develop more robust and adaptable models [77].

Furthermore, the integration of semi-supervised learning with continual learning, which addresses the challenge of learning from a stream of data with evolving distributions and tasks, holds great promise. In this setting, semi-supervised learning can play a crucial role in adapting to new data domains without forgetting previously learned knowledge, by leveraging unlabeled data to continuously expand the model's capabilities [40].

The emergence of large language models (LLMs) [25] has also opened up new possibilities for integrating semi-supervised learning. These powerful pre-trained models can serve as strong starting points for various downstream tasks, and the semi-supervised learning paradigm can be employed to further fine-tune and adapt these models to specific domains or tasks with limited labeled data.

The integration of semi-supervised learning with other machine learning paradigms, such as meta-learning and adversarial training, also holds promise. Meta-learning approaches can leverage the semi-supervised learning framework to rapidly adapt to new tasks or data distributions, while adversarial training can enhance the robustness and generalization of semi-supervised models [48].

Overall, the synergistic integration of semi-supervised learning with other machine learning paradigms can lead to the development of more robust, versatile, and data-efficient models that can thrive in challenging real-world scenarios. By capitalizing on the complementary strengths of various learning approaches, researchers can push the boundaries of what is possible with limited labeled data and pave the way for more practical and impactful applications of machine learning.

6.2 Advances in Robust and Scalable Semi-Supervised Algorithms

As the field of semi-supervised learning continues to advance, there is a growing emphasis on designing more robust and scalable algorithms that can handle the challenges encountered in real-world applications. Recent research has focused on developing techniques that can effectively address issues such as class imbalance, out-of-distribution samples, and noisy pseudo-labels, thereby improving the reliability and practicality of semi-supervised learning.

One prominent approach to address class imbalance in semi-supervised learning is the concept of Adaptive Dual-Threshold (ADT-SSL) [42]. Traditional semi-supervised methods often rely on a fixed threshold to determine which unlabeled samples to incorporate into the training process. However, this approach can lead to the exclusion of valuable information from unlabeled samples with lower confidence scores, which are typically harder samples and may belong to minority classes. ADT-SSL introduces an adaptive, class-specific threshold to capture these harder samples, effectively leveraging the information from a larger proportion of the unlabeled data. By engaging different loss functions for high-confidence and low-confidence unlabeled samples, ADT-SSL demonstrates superior performance on benchmark datasets compared to existing semi-supervised methods.

Another challenge in semi-supervised learning is the presence of out-of-distribution (OOD) samples in the unlabeled data, which can negatively impact the model's performance. To address this issue, researchers have proposed the Multi-Task Curriculum Framework for Open-Set Semi-Supervised Learning (MTCF-OSL) [78]. MTCF-OSL introduces a joint optimization framework that simultaneously learns to detect OOD samples and classifies in-distribution (ID) samples. By estimating the probability of a sample belonging to the OOD class, MTCF-OSL can selectively utilize the unlabeled ID samples for semi-supervised learning, while filtering out the OOD samples. This approach has shown promising results in handling the presence of OOD samples in the unlabeled data.

Another significant challenge in semi-supervised learning is the issue of noisy pseudo-labels, which can arise when the model's predictions on unlabeled samples are inaccurate. To address this, researchers have explored techniques that leverage ensemble-based approaches and self-supervised learning. For example, the Self-Training with Ensemble of Teacher Models (ST-ETM) [38] method utilizes an ensemble of teacher models to generate more reliable pseudo-labels for the unlabeled data, while also addressing the problem of model calibration. By carefully selecting the unlabeled samples to be included in the training process, ST-ETM demonstrates improved accuracy and calibration compared to traditional self-training approaches.

Additionally, the integration of self-supervised learning with semi-supervised learning has shown promising results in enhancing the robustness and scalability of semi-supervised algorithms. Techniques such as Contrastive Learning for Online Semi-Supervised General Continual Learning (SemiCon) [79] leverage self-supervised pretraining to learn robust representations from the unlabeled data, which can then be fine-tuned with limited labeled data. By leveraging the complementary strengths of self-supervised and semi-supervised learning, these hybrid approaches have achieved state-of-the-art performance on various benchmarks while demonstrating improved robustness to distribution shifts and class imbalance.

Furthermore, the exploration of energy-based models and adversarial training techniques in the semi-supervised learning context has also garnered significant attention. Energy-based models, such as those proposed in [58], can effectively utilize uncertainty estimates to identify and incorporate high-quality pseudo-labels from the unlabeled data, while rejecting low-confidence or out-of-distribution samples. Similarly, the integration of adversarial training, as demonstrated in [5], can enhance the model's robustness and generalization when leveraging both labeled and unlabeled data.

Overall, the recent advancements in robust and scalable semi-supervised learning algorithms have significantly improved the practicality and applicability of these techniques in real-world scenarios. By addressing challenges such as class imbalance, out-of-distribution samples, and noisy pseudo-labels, the research community has made substantial progress in developing semi-supervised learning methods that are more reliable, adaptable, and capable of effectively leveraging the abundance of unlabeled data available in many domains.

6.3 Applications of Semi-Supervised Learning in Emerging Domains

The application of semi-supervised learning in emerging domains has the potential to unlock new frontiers and drive impactful real-world applications. One such domain is medical imaging, where the scarcity of labeled data poses a significant challenge. Medical datasets, such as those used for disease diagnosis or image segmentation, often suffer from a lack of expert-annotated ground truth, as the process of manual labeling is time-consuming and expensive. Semi-supervised learning techniques can be leveraged to alleviate this burden by effectively utilizing the abundant unlabeled medical images [19; 58].

For instance, in the context of medical image segmentation, researchers have proposed semi-supervised methods that combine consistency regularization and pseudo-labeling to train deep neural networks [58]. By enforcing the model's predictions to be consistent under different perturbations of the input images and generating reliable pseudo-labels for the unlabeled data, these approaches have demonstrated significant performance improvements over fully supervised counterparts, especially when labeled data is scarce.

Another emerging domain where semi-supervised learning has shown promise is bioinformatics, particularly in tasks such as protein structure prediction and gene expression analysis. The inherent complexity and high dimensionality of biological data, coupled with the difficulty in obtaining labeled samples, make semi-supervised techniques an attractive solution. Researchers have explored semi-supervised approaches, such as graph-based methods and generative models, to leverage the wealth of unlabeled biological data and improve the performance of predictive models [80].

In the field of natural language processing (NLP), the emergence of large language models (LLMs) [63; 64] has opened up new opportunities for semi-supervised learning. LLMs, trained on vast amounts of unlabeled text data, have demonstrated remarkable few-shot learning capabilities, suggesting that the learned representations can be effectively fine-tuned or adapted to various downstream NLP tasks with limited labeled data. Semi-supervised techniques, such as self-training and consistency regularization, can be seamlessly integrated with LLM-based approaches to further boost performance in low-resource scenarios [40].

Moreover, semi-supervised learning can have a significant impact in domains where data collection and annotation are particularly challenging, such as environmental monitoring, robotics, and edge computing. In these settings, the ability to leverage unlabeled data can lead to more efficient and cost-effective model development, enabling the deployment of intelligent systems in resource-constrained environments.

Furthermore, the integration of semi-supervised learning with other machine learning paradigms, such as few-shot learning, continual learning, and meta-learning, can lead to powerful hybrid approaches that can adapt to evolving data distributions and handle tasks with scarce labeled data. By combining the strengths of these techniques, researchers can develop more robust and versatile models that can thrive in the face of real-world challenges [3].

In conclusion, the applications of semi-supervised learning in emerging domains, such as medical imaging, bioinformatics, and natural language processing, hold immense potential. By effectively leveraging unlabeled data and adapting to the unique characteristics of these domains, semi-supervised techniques can drive tangible real-world impact and pave the way for transformative advancements in various fields. As the field continues to evolve, the integration of semi-supervised learning with other cutting-edge machine learning approaches promises to yield even more powerful and adaptable solutions to address the challenges of the data-scarce era.

6.4 Open Challenges and Future Research Directions

The survey on semi-supervised learning has highlighted several key insights and the significant progress made in this field. However, there are still numerous open challenges and promising future research directions that warrant further investigation.

One critical challenge is the scalability of semi-supervised learning techniques to large-scale datasets. While the existing methods have shown promising results on standard benchmark datasets, their performance and computational efficiency may degrade when dealing with massive amounts of unlabeled data [21]. Developing scalable semi-supervised learning algorithms that can effectively leverage large-scale unlabeled data without compromising model performance is an important area for future research.

Another pressing challenge is the robustness of semi-supervised learning models to distribution shifts [5]. In real-world applications, the unlabeled data used during training may come from a different distribution than the test data, leading to significant performance degradation. Designing semi-supervised learning techniques that are resilient to distribution shifts, can handle out-of-distribution samples in the unlabeled data, and maintain consistent performance across diverse datasets is crucial for practical deployment.

The integration of semi-supervised learning with other machine learning paradigms also holds immense potential for future research. For instance, the combination of semi-supervised learning with few-shot learning [26] and continual learning could enable models to learn effectively from limited labeled data while adapting to new tasks and domains. Similarly, the synergy between semi-supervised learning and reinforcement learning or generative modeling could lead to advancements in areas such as decision-making, anomaly detection, and data synthesis.

Another promising direction is the theoretical understanding of semi-supervised learning. The survey has highlighted the need for a more comprehensive analysis of the assumptions, convergence guarantees, and generalization bounds associated with different semi-supervised learning techniques [3]. Deriving tighter theoretical bounds and understanding the fundamental limits of semi-supervised learning can guide the development of more principled and effective algorithms.

The impact of data quality and distribution on the performance of semi-supervised learning also requires further investigation. The survey has shown that the effectiveness of semi-supervised learning can be highly dependent on the characteristics of the labeled and unlabeled data, such as the degree of class imbalance, the presence of out-of-distribution samples, and the similarity between the labeled and unlabeled data distributions [81]. Exploring techniques to adaptively select or generate high-quality unlabeled data, and designing semi-supervised learning methods that are robust to various data distribution shifts, could lead to significant advancements in this field.

Finally, the successful application of semi-supervised learning in emerging domains, such as medical imaging, bioinformatics, and natural language processing, presents exciting opportunities for future research [68]. Adapting semi-supervised learning techniques to address the unique challenges and requirements of these domains, while leveraging the rich unlabeled data available, can have a profound impact on real-world applications.

In conclusion, the survey has highlighted the significant progress made in semi-supervised learning and the immense potential for further advancements. By addressing the open challenges related to scalability, robustness, integration with other learning paradigms, theoretical understanding, and domain-specific applications, the field of semi-supervised learning can continue to evolve and offer powerful solutions for a wide range of machine learning problems.

References

[1] A Review of Semi Supervised Learning Theories and Recent Advances

[2] A Survey on Semi-Supervised Learning Techniques

[3] Improvability Through Semi-Supervised Learning A Survey of Theoretical Results

[4] Asymptotic Bayes risk for Gaussian mixture in a semi-supervised setting

[5] Poisoning the Unlabeled Dataset of Semi-Supervised Learning

[6] Unsupervised Selective Labeling for More Effective Semi-Supervised Learning

[7] Semi-supervised Classification Cluster and Label Approach using Particle Swarm Optimization

[8] Patterns for Learning with Side Information

[9] The information-theoretic value of unlabeled data in semi-supervised learning

[10] Bayesian Semi-supervised learning under nonparanormality

[11] Graph-based Semi-supervised Learning A Comprehensive Review

[12] Augmenting Variational Autoencoders with Sparse Labels A Unified Framework for Unsupervised, Semi-(un)supervised, and Supervised Learning

[13] Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning

[14] Curriculum Labeling Revisiting Pseudo-Labeling for Semi-Supervised Learning

[15] Graph Laplacian for Semi-Supervised Learning

[16] A Modular Theory of Feature Learning

[17] Asymptotic Justification of Bandlimited Interpolation of Graph signals for Semi-Supervised Learning

[18] Random Matrix Analysis to Balance between Supervised and Unsupervised Learning under the Low Density Separation Assumption

[19] Semi-Supervised Learning with Deep Generative Models

[20] Deep Transductive Semi-supervised Maximum Margin Clustering

[21] Realistic Evaluation of Deep Semi-Supervised Learning Algorithms

[22] Fairness in Semi-supervised Learning Unlabeled Data Help to Reduce Discrimination

[23] Bridging the Gap Learning Pace Synchronization for Open-World Semi-Supervised Learning

[24] The Rich Get Richer Disparate Impact of Semi-Supervised Learning

[25] Rethinking Semi-supervised Learning with Language Models

[26] Semi-Unsupervised Learning Clustering and Classifying using Ultra-Sparse Labels

[27] Navigating the Pitfalls of Active Learning Evaluation A Systematic Framework for Meaningful Performance Assessment

[28] Split Batch Normalization Improving Semi-Supervised Learning under Domain Shift

[29] RealMix Towards Realistic Semi-Supervised Deep Learning Algorithms

[30] Consistent Semi-Supervised Graph Regularization for High Dimensional Data

[31] Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks

[32] Virtual Adversarial Training A Regularization Method for Supervised and Semi-Supervised Learning

[33] mixup Beyond Empirical Risk Minimization

[34] FixMatch Simplifying Semi-Supervised Learning with Consistency and Confidence

[35] Label Propagation for Deep Semi-supervised Learning

[36] Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning

[37] Rethinking Re-Sampling in Imbalanced Semi-Supervised Learning

[38] Self Training with Ensemble of Teacher Models

[39] SoftMatch Addressing the Quantity-Quality Trade-off in Semi-supervised Learning

[40] Revisiting Self-Training with Regularized Pseudo-Labeling for Tabular Data

[41] Confident Sinkhorn Allocation for Pseudo-Labeling

[42] ADT-SSL Adaptive Dual-Threshold for Semi-Supervised Learning

[43] Unsupervised Visual Representation Learning by Context Prediction

[44] Momentum Contrast for Unsupervised Visual Representation Learning

[45] Boosting the Performance of Semi-Supervised Learning with Unsupervised Clustering

[46] Semi-supervised Learning with Contrastive Predicative Coding

[47] Self-Tuning for Data-Efficient Deep Learning

[48] Generative Semi-supervised Learning with Meta-Optimized Synthetic Samples

[49] Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

[50] Meta-Semi A Meta-learning Approach for Semi-supervised Learning

[51] Adversarial Meta-Learning

[52] Variational Information Distillation for Knowledge Transfer

[53] Meta-Learning across Meta-Tasks for Few-Shot Learning

[54] Semi-supervised Contrastive Outlier removal for Pseudo Expectation Maximization (SCOPE)

[55] GEDI GEnerative and DIscriminative Training for Self-Supervised Learning

[56] Meta Pseudo Labels

[57] Unbiased Teacher for Semi-Supervised Object Detection

[58] Uncertainty-Aware Deep Co-training for Semi-supervised Medical Image Segmentation

[59] Reliable Semi-Supervised Learning when Labels are Missing at Random

[60] Semi-supervised learning method based on predefined evenly-distributed class centroids

[61] GuidedMix-Net Learning to Improve Pseudo Masks Using Labeled Images as Reference

[62] DiverseNet Decision Diversified Semi-supervised Semantic Segmentation Networks for Remote Sensing Imagery

[63] Language Models are Few-Shot Learners

[64] PaLM Scaling Language Modeling with Pathways

[65] Meta-learning of semi-supervised learning from tasks with heterogeneous attribute spaces

[66] Robust Deep Semi-Supervised Learning A Brief Introduction

[67] Semi-Supervised Deep Learning for Fully Convolutional Networks

[68] Systematic comparison of semi-supervised and self-supervised learning for medical image classification

[69] On semi-supervised learning

[70] Manifold regularization with GANs for semi-supervised learning

[71] Semi-Supervised Learning, Causality and the Conditional Cluster Assumption

[72] Distributed Representations of Words and Phrases and their Compositionality

[73] When can unlabeled data improve the learning rate

[74] Semi-supervised Data Representation via Affinity Graph Learning

[75] On the pitfalls of entropy-based uncertainty for multi-class semi-supervised segmentation

[76] Semi-Supervised Learning in the Few-Shot Zero-Shot Scenario

[77] ActiveMatch End-to-end Semi-supervised Active Representation Learning

[78] Multi-Task Curriculum Framework for Open-Set Semi-Supervised Learning

[79] Contrastive Learning for Online Semi-Supervised General Continual Learning

[80] Compressive Learning for Semi-Parametric Models

[81] Pruning the Unlabeled Data to Improve Semi-Supervised Learning