Data Optimization in Deep Learning: A Survey

Ou Wu, Rujing Yao

***Abstract*—Large-scale, high-quality data are considered an essential factor for the successful application of many deep learning techniques. Meanwhile, numerous real-world deep learning tasks still have to contend with the lack of sufficient amounts of high-quality data. Additionally, issues such as model robustness, fairness, and trustworthiness are also closely related to training data. Consequently, a huge number of studies in the existing literature have focused on the data aspect in deep learning tasks. Some typical data optimization techniques include data augmentation, logit perturbation, sample weighting, and data condensation. These techniques usually come from different deep learning divisions and their theoretical inspirations or heuristic motivations may seem unrelated to each other. This study aims to organize a wide range of existing data optimization methodologies for deep learning from the previous literature, and makes the effort to construct a comprehensive taxonomy for them. The constructed taxonomy considers the diversity of split dimensions, and deep sub-taxonomies are constructed for each dimension. On the basis of the taxonomy, connections among the extensive data optimization methods for deep learning are built in terms of five aspects. We probe into rendering several promising and interesting future directions. The constructed taxonomy and the revealed connections will enlighten the better understanding of existing methods and the design of novel data optimization techniques. Furthermore, our aspiration for this survey is to promote data optimization as an independent subdivision of deep learning. A curated, up-to-date list of resources related to data optimization in deep learning is available at https://github.com/YaoRujing/Data-Optimization.**

***Index Terms*—Deep learning, data optimization, data augmentation, sample weighting, data perturbation.**

I. Introduction

DEEP learning has received increasing attention in both the AI community and many application domains due to its superior performance in various machine-learning tasks in recent years. A successful application of deep learning cannot leave the main factors, which include a properly designed deep neural network (DNN), a set of high-quality training data, and a well-suited learning strategy (e.g., initialization schemes for hyper-parameters). Among the main factors, training data is of great importance and it usually plays a decisive role in the entire training process [1]. The concept of data-centric AI is rising, which breaks away from the widespread model-centric perspective [2]. Large models like GPT-4 show significant potential in the direction of achieving general artificial intelligence (AGI). It is widely accepted that the training for large models requires a huge size of high-quality training data.

Ou Wu is with Center for Applied Mathematics, Tianjin University, Tianjin, China, 300072, and with HIAS, University of Chinese Academy of Sciences, Hanzhou, China, 310000. Rujing Yao is with Department of Information Resources Management, Business School, Nankai University, Tianjin, China, 300071. E-mail: wuou@tju.edu.cn, rjyao@mail.nankai.edu.cn.

Manuscript received November 28, 2023. *(Corresponding author: Rujing Yao.)*

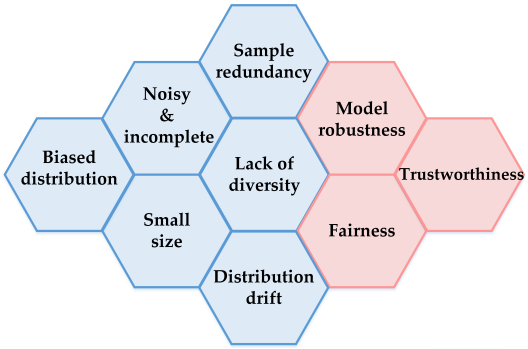

Fig. 1. Nine issues around real training data. The rightmost three issues focus on the DNN models.

However, most real applications lack ideal training data. Real training data usually encounters one or several of the nine common issues as shown in Fig. 1. The six issues on the left of Fig. 1 are directly related to training data:

(1) Biased distribution: This issue denotes that the distribution of the training data does not conform to the true distribution in a learning task. One typical bias is class imbalance, in which the proportions of different categories in the training data are not identical due to reasons such as data collection difficulties, whereas the proportions of different categories in test data are equal[1].

(2) Noisy and incomplete: This issue corresponds to at least two scenarios. The first refers to data noise that either partial training samples or partial training labels contain noises. As for sample noises, partial samples themselves are corrupted by noises. Taking optical character recognition (OCR) for example, some scanned images may contain serious background noises. The second occurs in multi-model/multi-view learning scenarios. Inconsistency and information missing may exist [3], [4]. For instance, the text title for an image may be mistakenly provided, or it may contain limited words.

(3) Small size: The training size surely impacts the training performance [5]. The larger the training data, the better the training performance usually being attained. Due to insufficient data collection budget or technique limitation, the training data will be relatively small for real use. Therefore, learning under small-size training data is a serious concern in deep learning. This study does not discuss the extreme cases of small size, such as few/one/zero-shot learning.

(4) Sample redundancy: Even though large training data is expected, it does not mean that every datum is useful.

[1]It should be noted that, in many imbalanced learning tasks, the distributions of the training data are unbiased and match those of the test data. However, the performance evaluation measures typically assume that the category distribution in the test data is uniform. Therefore, we also categorize these tasks under the biased distribution issue.

There are still learning tasks that the training set contains redundant data [6]. Two typical cases exist. First, the training size is relatively large and exceeds the processing capacity of the computing hardware. Second, some regions of training samples may be sampled excessively, and the deletion of such excessive samples does not affect the training performance. In this case, sample redundancy may occur in certain subsets of some categories.

(5) Lack of diversity: This issue refers to the fact that some attributes for certain categories concentrate excessively in the training corpus. Data diversity is also crucial for DNN training [7]. The lack of diversity in some non-essential attributes can lead to a spurious correlation between some non-essential attributes and the category. This issue is similar to the second case of sample redundancy. Nevertheless, lack of diversity does not necessarily imply the presence of redundant samples.

(6) Distribution drift: This issue denotes that the distribution of the involved data varies over time. Indeed, distribution drift may occur in most real learning applications, as either the concept or the form (e.g., object appearances, text styles) of samples varies fast or slow. Concept drift [8] is the research focus in distribution drift.

The above summary of data issues is not mutually exclusive, as there are overlaps among different issues. For example, small size may only occur in several categories, which can also be attributed to a type of biased distribution. Besides these data issues, there are also other (not exhaustive) model-related issues that are significantly influenced by the training data:

(7) Model robustness: This issue concerns the resistance ability of a DNN model to adversarial attacks [9]. If models for these applications are compromised by adversarial attacks, serious consequences may ensue.

(8) Fairness: This issue concerns the performance differences among different categories or attributes in a learning task [10]. For example, the recognition accuracy of faces in different color groups should be at the same level.

(9) Trustworthiness. This issue has emerged in many safety-critical AI applications [11]. It is closely related to robustness and fairness, and mainly refers to the explainability and calibration of DNN models.

To address the above-mentioned issues, numerous theoretical explorations have been conducted and tremendous new methodologies have been proposed in previous literature. Most of these existing methods directly optimize the involved data in learning rather than explore new DNN structures, which is referred to as **data optimization** for deep learning in this paper. As the listed issues belong to different machine learning divisions, the inspirations and focuses of these methods are usually distinct and seem unrelated to each other. For instance, the primary learning strategy for imbalanced learning (belonging to the biased distribution issue) is sample weighting which assigns different weights to training samples in deep learning training epochs. The primary manipulation for the small-size issue is to employ the data augmentation technique such as image resize and mixup [12] for image classification. When dealing with label noise in deep learning, one strategy is to identify noisy labels and then remove them during training. In cases where training data for certain categories lack sufficient diversity, causal learning is employed to break down the spurious correlations among labels and some irrelevant attributes such as certain backgrounds. Due to the apparent lack of connection, these studies typically do not mutually cite or discuss each other.

Our previous study [13] partially reveals that one technique, namely, data perturbation, has been leveraged to deal with most aforementioned issues. This observation illuminates us to explore the data optimization methodologies for those issues in a more broad view. In this study, a comprehensive review for a wide range of data optimization methods is conducted. First, a systematic data optimization taxonomy is established in terms of eight dimensions, including pipeline, object, technical path, and so on. Second, the intrinsic connections among some classical methods are explored from five aspects, including data perception, application scenarios, similarity/opposition, theory, and data types. Third, theoretical studies are summarized for the existing data optimization techniques. Lastly, several future directions are presented according to our analysis.

The differences between our survey and existing surveys in relevant areas, including imbalanced learning, noisy-label learning, data augmentation, adversarial training, and dataset distillation, lie in two aspects. First, this survey takes a data-centric view for studies from a wide range of distinct deep learning realms. Therefore, our focus is merely on the data optimization studies for the listed issues. Methods that do not belong to data optimization for the listed issues are not referred to in this study. Second, the split dimensions (e.g., data perception and theory) which facilitate the establishment of connections among seemingly unrelated methods are considered in our taxonomy. These dimensions are usually not referred to in the existing surveys.

The contributions of this study are summarized as follows.

- Methodologies related to data enhancement for dealing with distinct deep learning issues are reviewed with a new taxonomy. To our knowledge, this is the first work that aims to construct a data-centric taxonomy focusing on data optimization across multiple deep learning divisions.
- The connections among many seemingly unrelated methods are built according to our constructed taxonomy. The connections can inspire researchers to design more potential new techniques.
- Theoretical studies for data optimization are summarized and interesting future directions are discussed.

This paper is organized as follows. Section II introduces relevant survey studies. Section III describes the main framework of our constructed taxonomy. Sections IV, V, VI, and VII introduce the details of our taxonomy. Section VIII explores the connections among different data optimization techniques. Section IX presents several future directions, and conclusions are presented in Section X.

II. Related studies

The issues listed in the previous section gradually spawn numerous independent research realms of deep learning. Subsequently, there have been many survey studies conducted

for these issues. The following introduces related surveys in several typical research topics.

Imbalanced learning. It is a hot research area in deep learning [14]. He and Garcia [15] conducted the first comprehensive yet deep survey study on imbalanced learning. They explored the intrinsic characteristics of learning tasks incurred by imbalanced data. Recent studies have focused on the extreme case of imbalanced learning, namely, long-tailed classification. Zhang et al. [16] summarized the recent developments in deep long-tailed classification. In their constructed taxonomy, module improvement such as a new classifier is listed as one of the three main techniques. In this study, module improvement is not considered, as it does not fall under data optimization.

Noisy-label learning. This is another research area gaining tremendous attention as label noise is nearly unavoidable in real applications. Algan and Ulisory [17] summarized the methods in noisy-label learning for image classification. Song et al. [18] elaborately designed taxonomy for noisy-label learning along with three categories, including "data", "objective", and "optimization".

Learning with small data. Big data has achieved great success in deep learning tasks. Meanwhile, many real learning tasks still confront with the challenge of small-size training data. Cao et al. [19] performed rigorous theoretical analysis for the generalization error and label complexity of learning on small data. Wang et al. [20] constructed a few-shot learning taxonomy with three folds, including "data", "model", and "algorithm". Data-centric learning methods are also among the primary choices for few-shot learning.

Concept drift. Lu et al. [8] investigated the learning for concept drift under three components, including concept drift detection, concept drift understanding, and concept drift adaptation. Yuan et al. [21] divided existing studies into two categories, namely, model parameter updating and model structure updating in concept drift adaptation. This division is from the viewpoint of the model. Indeed, pure data-based strategy, such as data augmentation [22], is also employed in learning under concept drift.

Adversarial robustness. In many studies, model robustness is limited to adversarial robustness. Silva and Najafirad [23] divided adversarial robust learning methods into three categories, including adversarial training, regularization, and certified defenses. Xu et al. [24] summarized the studies for model robustness on graphs. Goyal et al. [25] reviewed the adversarial defense and robustness in natural language processing.

Fairness-aware learning. It receives increasingly attention in recent years. Mehrabi et al. [26] explored different sources of biases that can affect the fairness of learning models. They revealed that each of the three factors, namely, data, learning algorithms, and involved users may result in bias. Sample reweighting and adversarial training are two common strategies for fair machine learning [27].

Trustworthy learning. It is the key of trustworthy AI, which aims to ensure that an AI system is worthy of being trusted. Trust is a complex phenomenon [28] highly related to fairness, explainability, reliability, etc. Kaur et al. [29] summarized studies on trustworthy artificial intelligence in a quite broad view. Wu et al. [30] provided an in-depth review for studies about trustworthy learning on graphs.

There are also studies that focus on learning tasks with more than one of the listed data issues. For example, Fang et al. [31] addressed noisy-label learning under the long-tailed distributions of training data. Singh et al. [32] conducted an empirical study concerning fairness, adversarial robustness, and concept drift, simultaneously. To our knowledge, no survey study pays attention to the intersection of the research areas related to the listed issues. The unified taxonomy constructed in this survey will enlighten the study on the intersection of multiple areas.

The most similar study to this work is the survey presented by Wan et al. [33], which focuses on data optimization in computer vision. There are significant differences between our and Wan et al.'s study. First, the covered technical scopes of ours are much broader than those of Wan et al.'s study. Their study limits the scope merely in data selection, including resampling, subset selection, and active learning-based selection. Second, the split dimensions of ours are quite different from those in [33] for the overlapped methods. Lastly, additional important parts including data perception, connections, and theoretical investigation are introduced and discussed in this study.

The topic investigated in this study falls under data-centric AI [34] and, more specifically, its subdivision, data-centric deep learning [35]. Zha et al. [36] provided a clear, high-level summary of recent studies on data-centric AI. In their data-centric AI taxonomy, there are three main components: training data development, inference data development, and data maintenance. Our study primarily focuses on training data development. However, there are two differences between Zha et al.'s summary on training data development and this survey. First, Zha et al. categorized training data development methods into five categories: collection, labeling, preparation, reduction, and augmentation. In contrast, the first three categories are not explicitly addressed, and a broader range of data processing aspects is covered in this study. Second, Zha et al.'s study presents a high-level overview of data development, whereas this study provides a fine-grained description of each data optimization technique. There are also surveys on data-centric engineering [37], [38]. Pan et al. [38] surveys data-centric studies in the context of chemical engineering. There are two main differences between these studies and our study. First, these data-centric engineering surveys organize existing methodologies according to the data processing pipeline, meaning that machine learning-related data processing is only one part of these studies, whereas our study primarily focuses on deep learning. Second, a significant portion of these studies addresses specific data processing components within their respective engineering fields, while our study discusses more general application contexts.

Traditional shallow machine learning also heavily relies on the quality of training data, making data optimization a widely explored area in shallow learning. However, data optimization for shallow learning primarily focuses on data cleaning, resampling, and weighting. In contrast, deep learning, which requires large training datasets and is more sensitive to computational complexity, favors data augmentation and data pruning. Nevertheless, numerous optimization methods

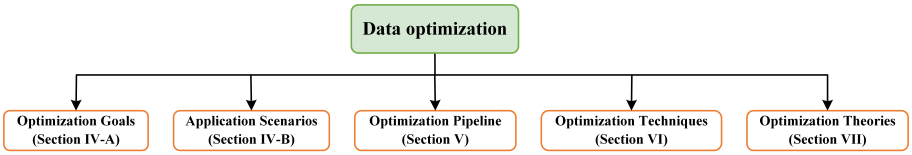

Fig. 2. The five split dimensions of our constructed taxonomy for data optimization.

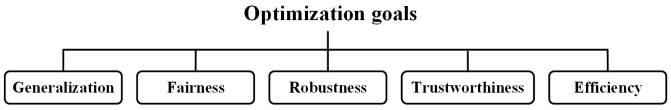

Fig. 3. The sub-taxonomy for optimization goals.

and principles from shallow learning have been adapted for deep learning. Many resampling and weighting methods in deep learning are directly inspired by or derived from those used in shallow learning tasks. Classical data augmentation techniques in shallow learning, such as SMOTE [39], are also adapted to deep learning tasks.

III. Overall of The Proposed Taxonomy

To ensure our constructed taxonomy well organized and comprehensive coverage on previous data optimization techniques about the issues listed in Section I as much as possible, the following principles for the split dimensions are adopted.

(1) The first layer of the taxonomy should consider multiple views, with each view corresponding to a sub-taxonomy. Most existing taxonomies for specific research realms adopt only a single view. In this study, only a single view is inadequate for systematically arranging studies from various deep learning realms.
(2) The dividing dimension should be general so as to embrace existing data optimization studies as much as possible. Therefore, the dimensions designed in existing taxonomies for specific research areas should not be directly followed. A new comprehensive taxonomy is required.
(3) The new taxonomy should be compatible with existing taxonomies. That is, inconsistency between our and existing taxonomies is allowed. However, contradiction between them should be avoided.

On the basis of these principles, the first layer[2] of our taxonomy is designed as shown in Fig. 2. This layer consists of five dimensions for data optimization as follows:

- Optimization goals. This dimension refers to the final goal of a data optimization method used in a deep learning task. We divide the optimization goals into five important aspects[3], including generalization, robustness, fairness, trustworthy, and efficiency.
- Application scenarios. This dimension refers to the deep learning applications that utilize data optimization. Nine applications are involved, including learning under biased distribution, noisy-label learning, learning with redundant training data, Safety-aware learning, fairness-aware learning, learning under distribution drift, trustworthy learning, learning under insufficient data, and learning for large models.
- Optimization pipeline. This dimension refers to the common steps for data optimization. There are three common steps, namely, perception, analysis, and optimizing.
- Optimization techniques. This dimension refers to the technical paths in data optimization. This study summarizes five main technical paths. Each path contains a sub-division. This part is the focus of this survey.
- Optimization theories. This dimension refers to the theoretical analysis and exploration for data optimization in deep learning. We divided this dimension into two aspects, namely, formalization and explanation.

Section IV introduces the ultimate goals and application scenarios. Sections V, VI, and VII introduces the optimization pipeline, techniques, and theories, respectively.

IV. Goals and Scenarios

A. Optimization goals

Fig. 3 describes the sub-taxonomy for the dimension of optimization goals, including generalization, fairness, robustness, trustworthiness, and efficiency.

Generalization is the primary optimization goal in most data optimization techniques, as it is almost the sole goal in most deep learning tasks. According to the generalization theory studied in shallow learning, generalization of a category is highly related to class margin, inter-class distance, and class compactness [40]. The data augmentation strategy that injects noise to training samples is proven to increase the generalization [41]. The implicit data augmentation method ISDA [42] actually improves each category's class compactness[4]. Adaptive margin loss [43] also aims to improve the class compactness by perturbing the logits. Fujii et al. [44] modified the augmentation method mixup [12] by considering the "between-class distance" to increase the inter-class distance.

As previously stated, fairness is also an important learning goal in many deep learning tasks. To combat unfairness on samples with certain attributes, techniques such as data augmentation [45], perturbation [46], and sample weighting [47] have been used in previous literature.

Adversarial robustness is an essential goal in deep learning tasks that are quite sensitive to model safety [48]. Adversarial training is usually leveraged to improve the adversarial robustness of a model. It can be attributed to a special type of data augmentation [49].

Trustworthiness is a goal that has recently been highly valued. Explainability and calibration are its two crucial requirements. Data optimization, such as perturbation [50] and weighting [51], is widely used in model calibration. Calibration mainly concerns the trustworthiness of the predicted probability of a probabilistic model [52].

[2]The fine-granularity layers are detailed in the succeeding sections.

[3]It should be noted that these five aspects are not exhaustive and there are overlaps among them as revealed by the previous literature.

[4]Some methods such as center loss also aim to increase the class compactness. These methods are considered not data optimization.

Fig. 4. The sub-taxonomy for targeted application scenarios.

Efficiency is crucial for real applications as many learning tasks are sensitive to both time complexity and storage. Therefore, how to optimally reduce the redundant training data and remain the diverse important training data deserves further investigation [53]. The time complexity can be significantly reduced after data pruning.

B. Application scenarios

Fig. 4 describes the sub-taxonomy for the dimension of targeted application scenarios. The first seven scenarios have been referred to in previous sections, so they are not further introduced in this subsection. Here only discusses learning under insufficient data and learning for large models.

Learning under insufficient data contains the case that the training data are not as diverse as possible. Data diversity affects the model generalization [7]. Dunlap et al. [54] utilized large vision and language models to automatically generate visually consistent yet significantly diversified training data. Some studies [55] consider that data augmentation actually increases data diversity.

Large models have made remarkable advancements recently. The data quality is crucial for the training of a large model. Yang et al. [56] utilized flip operation on the training corpus to balance the two-way translation in language pairs in their building of a large model. Liu et al. [57] applied adversarial training in both the pre-training and fine-tuning.

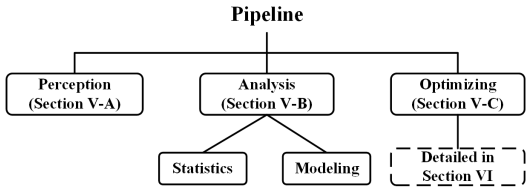

Fig. 5. Three main steps in data optimization pipeline.

V. Optimization Pipeline

The pipeline mainly consists of three steps, namely, perception, analysis, and optimizing, as shown in Fig. 5.

A. Data perception

In this study, data perception refers to all possible methods aimed at sensing and diagnosing the training data to capture the intrinsic data characteristics and patterns that affect learning performance. It serves as the first step in the pipeline, and an effective data optimization method cannot work well without accurate perception of the training data.

Generally, data perception for training data quantifies the factors related to the true distribution, training data distribution, cleanliness, diversity, etc. We construct a sub-taxonomy for data perception in three dimensions as shown in Fig. 6.

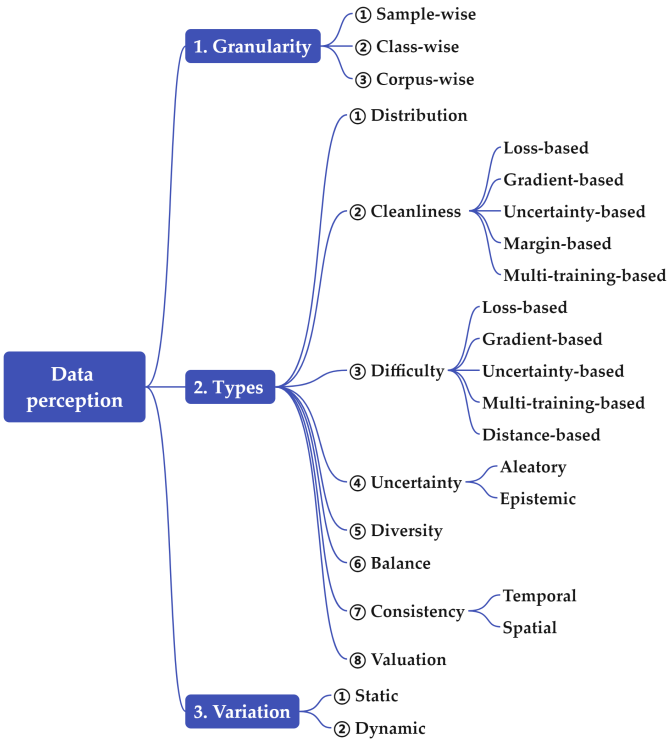

Fig. 6. The sub-taxonomy for data perception.

First, in terms of quantifying granularity, there are three levels, namely, sample-wise, category-wise, and corpus-wise. Secondly, in terms of perception types, there are eight divisions, namely, distribution, cleanliness, difficulty, uncertainty, diversity, balance, consistency, and valuation. Thirdly, in terms of quantifying variation, there are two divisions, namely, static and dynamic methods. Each of the above divisions is introduced as follows.

1) Perception on different granularity levels: There are three granularity levels, including sample-wise, category-wise, and corpus-wise.

- Sample-wise data perception. It denotes that the perceived quantities reflect or influence a sample's positive/negative or trivial/important role in training. For example, training loss [58] and gradient norm [59] are widely used to infer the noisy degree of a training sample.
- Category-wise data perception. It denotes that the perceived quantities reflect or influence a category's positive/negative or trivial/important role in training. In category-wise perception, the learning performance [60], the proportion, or the compactness [61] of each category, are usually used monitored to return feedback for the entire scheme.
- Corpus-wise data perception. It denotes that the perceived quantities reflect or influence a training corpus' positive/negative or trivial/important role in training. Lin et al. [62] used the query score to measure the utility of a training dataset.

2) Perception on different types: The eight quantifying types are introduced as follows:

- Distribution. This type aims to quantify the true data distribution for a learning task and the training data distribution.

The true distribution is usually assumed to conform to several some basic assumptions, such as Gaussian distribution for each category [42]. For the training data distribution, some studies [63] apply clustering to deduce the intrinsic structure of the training data. Recently, researchers have investigated local distributions of training samples. One typical characteristic is about the neighborhood of each training sample [64]. Wang et al. [65] defined a label difference index to quantify the difference between a node and its neighborhood in a graph.

- Cleanliness. This type aims to identify the degree of noise in each sample. This study primarily focuses on label noise, as it garners more attention than sample noise. As illustrated in Fig. 6, there are five typical label noise measures, including loss-based, gradient-based, uncertainty-based, margin-based, and multi-training-based techniques. Samples with large losses, large gradient norms, large uncertainties, or small margins are likely to be noisy. Huang et al. [58] conducted multiple training procedures to identify noisy labels.
- Difficulty. This type aims to infer the degree of learning difficulty for a training sample or a category. The accurate measurement of learning difficulty for each training sample is of great importance because several deep learning paradigms employ adaptive learning strategies based on the level of learning difficulty, such as curriculum learning [66] and Focal loss [67]. As shown in Fig. 6, there are five major manners to measure learning difficulty of samples, namely, loss-based, gradient-based, uncertainty-based, multi-training-based, and distance-based. Obviously, the measures for learning difficulty are quite similar to those for cleanliness. In fact, some studies consider that noisy samples are those quite difficult to learn and divide samples into easy/medium/hard/noisy. Zhu et al. [68] established a formal definition for learning difficulty of samples inspired by the bias-variance trade-off theorem and proposed a new learning difficulty measures.
- Uncertainty. This type contains two sub-types, namely, aleatory uncertainty and epistemic uncertainty [69]. The former is also called data uncertainty and occurs when training samples are imperfect, e.g., noisy. Therefore, the cleanliness degree can be used as a measure of data uncertainty. Epistemic uncertainty is also called model uncertainty. It appears when the learning strategy is imperfect and can be calculated based on information entropy of the prediction [70].
- Diversity. This type aims to identify the diversity of a subset of training samples. The subset is usually a category. The measurement for subset diversity is useful in the design of data augmentation strategy for the subset [71] and data selection [72]. Friedman and Dieng [73] leveraged the exponential of the Shannon entropy of the eigenvalues of a similarity matrix, namely, vendi score to measure diversity. Pang et al. [74] designed a novel and efficient diversity measure, named instance Euclidean distance metric (IED), to evaluate diversity of a training subset.
- Balance. This type aims to measure the balance between/within categories. The balance between categories belongs to global balance, while that within a category belongs to local balance. Global balance can be simply measured by the proportion of the training sample of a category. Nevertheless, our previous study [75] reveals that other factors such as variance and distance may also result in serious imbalance.
- Consistency. This type aims to identify the consistency of the training dynamics of a training sample along the temporal or spatial dimensions. In the temporal dimension, the variations of the training dynamics between the previous and the current epochs are recorded [76]. In the spatial dimension, the differences in the training dynamics between a sample and other samples such as neighbors [64] or samples within the same category are recorded. A classical measure called "forgetting" quantifies the number of variations in the prediction between adjacent epochs. Wang et al. [77] provided a comprehensive summary for sample forgetting.
- Valuation. This value is usually measured by the Shapley value, which is a concept from the game theory [78]. Ghorbani and Zou firstly introduced Shapley value for data valuation [79]. Nevertheless, the calculation for the Shapley value is NP-hard, thereby hindering its use in real applications. Jiang et al. [80] established an easy-to-use and unified framework that facilitates researchers and practitioners to apply and compare existing algorithms.

This study only lists commonly used measures for data perception. Some other important quantities such as problematic score [81] and data influence [82] in learning, which have large overlaps with the aforementioned quantities, also deserve further exploration.

3) Static and dynamic perception: Static perception denotes that the perceived quantities remain unchanged during optimization, whereas dynamic perception denotes that the quantities vary.

In imbalanced learning, category proportion is widely used to quantify a category. It belongs to static perception because this quantity remains unchanged. In noisy-label learning, many studies adopt a two-stage strategy in which the noisy degree of each training sample is measured and the degrees are used in the second training stage [58]. In this two-stage strategy, the perception for label noise is static.

The impact of a training sample usually varies during training. Therefore, compared with static perception, dynamic perception is more prevailing in deep learning tasks. Many studies utilize training dynamics of training samples for the successive sample weighting or perturbation. Such training dynamics also belong to the dynamic perception. The training dynamics including loss, prediction, uncertainty, margin, and neighborhood vary at each epoch. For example, self-paced learning [83] determines the weights of each training sample according to their losses in the previous epoch and a varied threshold. Therefore, the weights may also vary in each epoch.

B. Analysis on perceived quantities

Analysis on perceived quantities contains two manners, namely, statistics and modeling, as shown in Fig. 5.

1) Statistical analysis: . Most studies employ this manner for the perceived data quantities. These studies considered only one or two quantities. For example, Toneva et al. [84] made

a statistics for the forgetting numbers of training samples and revealed that distinct difference exists between the distributions of clean and noisy samples. Huang et al. [58] proposed a new strategy that the model is trained from overfitting to underfitting cyclically. The epoch-wise loss for each training sample is recorded. Noisy samples have larger training losses. Therefore, they leveraged the average loss as an indicator for noisy labels. Zhu et al. [68] proposed a cross validation-based training strategy. Multiple training losses are also recorded for each training sample. They revealed that the variance of the multiple losses for each sample is also useful in identifying noisy labels.

2) Modeling: . This manner refers to the statistical modeling on the perceived quantities for training data. Arazo et al. [85] assumed that the training loss conforms to the Gamma distribution. Their values are different when modeling the clean and the noisy samples. Hu et al. [86] leveraged the Weibull mixture distribution to model the memorization-forgetting value of each sample, which can distinguish clean and noisy samples.

These two divisions usually rely on appropriate prior distributions about the involved quantities. If the prior distributions are incorrect, the successive optimizing will negatively influence the model training.

C. Optimizing

The data perception and analysis act as the pre-processing for data operation. This step is the key processing of the entire data optimization pipeline. The successive section will introduce current optimization techniques in detail.

VI. Data optimization techniques

This section describes the most important dimension for the presented taxonomy, namely, data optimization techniques for deep learning. Fig. 7 presents the sub-taxonomy along this dimension. We summarized six sub-divisions for existing data optimization techniques, including resampling, augmentation, perturbation, weighting, pruning, and others. It is noteworthy that this survey covers numerous technique/methodology divisions and leaves a through comparison for them as our future work. The reason lies in two folds. First, each division has its own merits and defects and their effectiveness have been verified in previous literature, so it is difficulty to judge which one is absolutely the best in arbitrary tasks. Second, a thorough theoretical or empirical comparison is not a trivial task.

A. Data resampling

Data resampling compiles a new training set in which training data are randomly sampled from the raw training set. It is widely used in tasks encountering the issues, including biased distribution [87] and redundancy. This study summarizes two split dimensions for this division. The first dimension concerns the size of the sampled dataset, while the second dimension concerns the sampling rate.

In the first dimension, resampling is divided into undersampling and oversampling. Undersampling compiles a new training set whose size is smaller than that of the raw training set. Contrarily, oversampling compiles a new training set whose size is larger than that of the raw training set. Both manners are widely used in previous learning tasks, including imbalanced learning, bagging, and cost-sensitive learning. Meanwhile, tremendous theoretical studies have been conducted to explain the effectiveness of these two manners in both the statistics and the machine learning communities. Nevertheless, there is currently no consensus on which manner is more effective. Some studies concluded that undersampling should be the primary choice when dealing with imbalanced datasets [88]. However, some other studies hold the opposite view [89].

In the second dimension, resampling is divided into the following five folds:

- Uniform sampling. This manner is quite intuitive. It treats samples definitely equal regardless of their distributions, location, categories, and training performances. In nearly all existing deep learning tasks, the batch is constructed by uniformly sampling from the training corpus. Some studies explore alternative sampling strategies. For example, Loshchilov and Hutter [90] proposed a rank-based batch selection strategy in which samples with large losses have high probabilities to be sampled.
- Proportion-based sampling. This manner simply assigns the total sampling rate for each category with its proportion (π_c) in the training corpus. It is mainly used in imbalanced learning in which the minor categories are assigned with large sampling rates [15].
- Importance-based sampling. This manner assigns sampling probabilities according to samples' importance. In this study, the definition for importance sampling follows several classical studies [91]. Given a target distribution $q(x, y)$ and a source distribution on training data $p(x, y)$, the importance (sampling rate) for a training sample $\{x, y\}$ in importance sampling is defined as

$$w(x) = \frac{q(x, y)}{p(x, y)}. \tag{1}$$

  As the target distribution is unknown, some studies [92] utilize the kernel trick to generate sampling rates. In some importance sampling studies, the sampling rates are not based on Eq. (1). For example, Atharopoulos and Fleuret [93] took the gradient norms of each training sample as their importance. This method actually belongs to learning difficulty-based sampling.
- Learning difficulty-based sampling. This manner assigns sampling rates according to samples' learning difficulties. As summarized in Section V-A2, learning difficulty is usually measured by loss or gradient norm. Johnson and Guestrin [94] proposed the O-SGD sampling method with the following sampling rate:

$$w(x, y) = \frac{||\nabla l(x, y)||}{\sum_x ||\nabla l(x, y)||}, \tag{2}$$

  where $l(x, y)$ is the training loss. They claimed that this "importance sampling" can reduce the stochastic gradient's variance and thus accelerate the training speed. Gui et al. [95] utilized sampling strategy for noisy-label learning. They calculated the sampling weights based on the mean

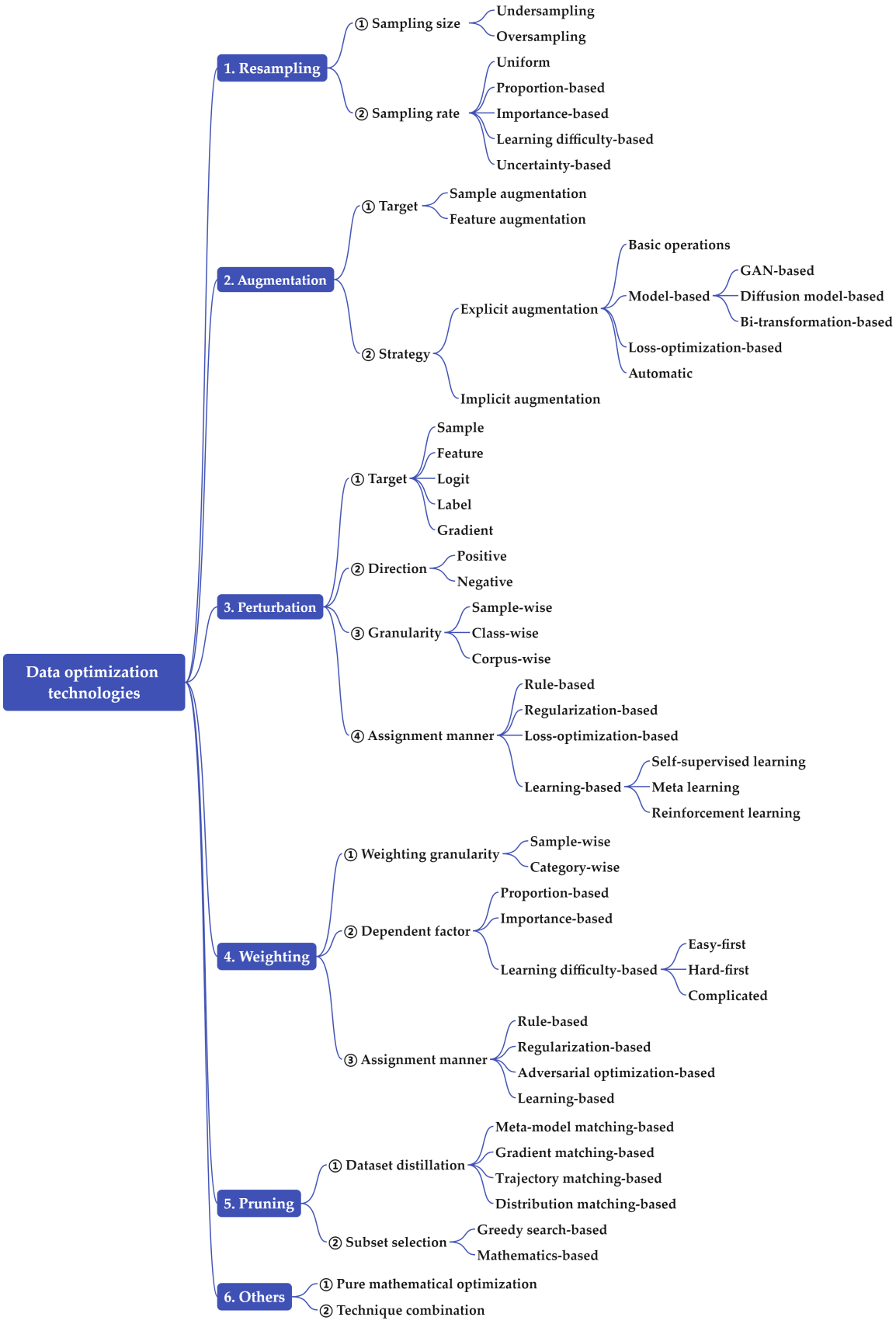

Fig. 7. The sub-taxonomy of data optimization techniques.

loss of each example along the training process. The training samples with large mean losses are assigned low weights.

- Uncertainty-based sampling. This manner assigns sampling rates according to samples' uncertainties. It is widely used in active learning, in which a subset of data is sampled for human labeling [96], [97]. Aljuhani et al. [98] presented an uncertainty-aware sampling framework for robust histopathology image analysis. The uncertainty is calculated by predictive entropy.

There are also some other sampling manners. For instance, Ting and Brochu [99] calculated the sample influence for optimal data sampling. Li and Vasconcelos [100] proposed the adversarial sampling to improve OOD detection performance of an image classifier. Zhang et al. [101] sampled training data of the majority categories by considering the samples' sensitivities. Sun et al. [102] explored an automatic scheme for effective data resampling.

B. Data augmentation

Data augmentation compiles a new training set in which samples (or features) are generated based on the raw training set or sometimes other relevant sets. It is a powerful tool to improve the generalization capability and even adversarial robustness of DNNs. Illuminated by related surveys on data augmentation [103], [104], two split dimensions are considered, namely, target (sample/feature) and strategy (explicit/implicit), as shown in Fig. 7.

1) Augmentation target: In sample augmentation, the new training set consists of generated new samples, while in feature augmentation, the new training set consists of generated new features.

- Sample augmentation. This division is subject to data types (e.g., image, text, or others). For image corpus, augmentation methods adopt noise adding, color transformation, geometric transformation, or other basic operations such as cropping to augment new images [104]. For texts, new samples can be generated by noise adding, paraphrasing, or other basic operations such as word swapping [105].
- Feature augmentation. Different from sample augmentation which is performed directly on raw samples, this division is performed on the extracted/transformed features from raw samples, so learning tasks for different data types may utilize the same or similar augmentation strategies. Some intuitive feature augmentation methods include adding noise, interpolating, or extrapolating, which are applicable for general data types, including both image and text data. Li et al. [106] revealed that the simply perturbing the feature embedding with Gaussian noise in training leads to comparable performance compared with the SOTA methods. Cui et al. [107] generated samples by combining two components, namely, class-generic and class-specific, for minor categories. Adversarial training, is actually a feature-wise augmentation strategy when it is run on the feature space.

Some studies augment other data targets such as label [108] and gradient [109], which receives quite limited attention.

2) Augmentation strategy: Explicit augmentation directly generates new samples/features. Meanwhile, implicit augmentation conducts data augmentation only theoretically, yet it does not generate any new samples/features actually.

- Explicit augmentation. As described in Fig. 7, explicit augmentation is divided into the following four folds:
  - Basic operations. This technique is widely used in practical learning tasks as basic operations conform to human intuitions. The popular deep learning platforms such as pyTorch provide several common basic operations such as cropping, rotation, replacement, masking, cutout, etc. One of the most popular data augmentation method used for shallow learning tasks, namely, SMOTE [39] has been utilized in deep learning tasks [110]. SMOTE generates new samples by using a linear combination of a selected sample and its neighbors. Inspired by SMOTE, Xie et al. [111] proposed a novel instance generation method to address imbalanced learning tasks. Among the basic operations, mixup is a simple yet quite effective manner [12]. The original Mixup algorithm generates a new sample based on the linear combination of two randomly selected training samples, similar to SMOTE. Nevertheless, it synthesizes a new label that does not belong to the raw label space.
  - Model-based augmentation. This technique generates new samples by leveraging independent models. There are three main schemes:
    ① GAN-based scheme. The classical generative adversarial network (GAN) is an unsupervised learning method [112]. The basic idea of GAN is a two-player zero-sum game: one player is a generative model attempting to generate fake data that closely resembles real data, while the other is a discriminative model trying to distinguish between generated data and real data. The two models can train simultaneously in a well designed two-player min-max game. The trained generative model can be used to generate new samples conforming to the distribution of the training data. Yang et al. [113] utilized the GAN-based augmentation for time series.
    ② Diffusion model-based scheme. Diffusion models are a new class of generative models and achieve SOTA performance in many applications [114]. They are deep learning generative models based on probabilistic statistics and the principles of non-equilibrium thermodynamics. Initially inspired by the molecular diffusion process in physics, these models are primarily used to learn the probability distribution of data and generate new samples. Dunlap et al. [54] utilized large vision and language models to automatically generate natural language descriptions of a dataset's domains and augment the training data via language-guided image editing.
    ③ Bi-transformation-based scheme. This scheme usually relies on two transformation models. The first model transforms a training sample into a new type of data. The second model transforms the new type of data into a new sample. In natural language processing, back-translation is a popular data augmentation tech-

nique [115], which translates the raw text sample into new texts in another language and back translates the new texts into a new sample in the same language with the raw sample.

– Loss-optimization-based augmentation. This manner generates new sample/features by minimizing or maximizing a defined loss with heuristic or theoretical inspirations. Adversarial training is a typical loss-optimization-based manner. It generates a new sample for $\boldsymbol{x}$ by solving the following maximization problem:

$$\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \arg\max_{\|\boldsymbol{\delta}\| \leq \epsilon} \ell(f(\boldsymbol{x} + \boldsymbol{\delta}), y), \tag{3}$$

where $\boldsymbol{\delta}$ and ϵ are the perturbation term and bound, respectively. Zhou et al. [49] proposed anti-adversaries by minimizing the loss. Pagliardini et al. [116] obtained new samples by maximizing an uncertainty-based loss.

– Automatic augmentation. This manner investigates automated data augmentation techniques based on meta learning [117] or reinforcement learning [118].

- Implicit augmentation. Wang et al. [42] proposed the first implicit augmentation method called ISDA. It establishes a Gaussian distribution for each category. New samples can be generated (i.e., sampled) from its corresponding distribution. An upper bound of the loss with augmented samples can then be derived when the number of generated samples for each training sample approaches to infinity. Finally, the upper bound of the loss is used in the training. ISDA has some variations, such as IRDA [119] and ICDA [120].

Explicit augmentation is the primary choice in data augmentation tasks. Nevertheless, implicit augmentation is usually more efficient than explicit augmentation as it does not actually generate any new samples or features.

C. Data perturbation

Given a datum x (x can be the raw sample, feature, logit, label, or others), data perturbation will generate a perturbation $\triangle x$ such that $x' = x + \triangle x$ can replace x or be used as a new datum. Therefore, some data augmentation methods, such as adversarial perturbation and masking, can also be viewed as data perturbation. In our previous work [13], we constructed a taxonomy for learning with perturbation. This study follows our previous taxonomy in [13] with slight improvements. The sub-taxonomy for data perturbation is presented in Fig. 7. The following four split dimensions are considered.

1) Perturbation target: The perturbation targets can be raw sample, feature, logit vector, label, and gradient.

- Sample perturbation. This division adds the perturbation directly to the raw samples. The basic operations in data augmentation can be placed into this division. For instance, noise addition and masking used in image classification actually exert a small perturbation on the raw image.
- Feature perturbation. This division adds the perturbation on the hidden features. Jeddi et al. [121] perturbed the features at each layer to increase uncertainty of the network. Their perturbation conforms to the Gaussian distribution. Shu et al. [122] designed a single network layer that can generate worst-case feature perturbations in training to improve the robustness of DNNs.
- Logit perturbation. This division adds the perturbation on the logit vectors in the involved DNNs. Li et al. [123] analyzed several classical learning methods such as logit adjustment [124] and ISDA [42] in a unified logit perturbation viewpoint. They proposed a new logit perturbation method for multi-label learning [125].
- Label perturbation. This division adds the perturbation on either the ground-truth label or the predicted label. One classical learning skill, namely, label smoothing [126], is a kind of label perturbation method. Wang et al. [127] proposed reward perturbation for noisy reinforcement learning.
- Gradient perturbation. This division adds the perturbation directly on gradient. Studies on gradient perturbation are few. Orvieto et al. [128] proposed a gradient perturbation method and verified its effectiveness theoretically.

There are also studies [129] which perturb other data such as network weights, which is not the focus of this study.

2) Perturbation direction: Data perturbation will either increase or decrease the loss values of training samples in the learning process. Based on whether the loss increases or decreases, existing methods can be categorized as positive or negative augmentations.

- Positive perturbation. If the perturbed training samples have larger losses than their raw samples, the corresponding method belongs to positive perturbation. Obviously, adversarial perturbation belongs to positive perturbation, as it maximizes the loss with the adversarial perturbations. ISDA [42] also belongs to positive perturbation as it adds positive real numbers to the denominator of the Softmax.
- Negative perturbation. If the perturbed training samples have smaller losses than their raw samples, the corresponding method belongs to negative perturbation. Anti-adversarial perturbation [49] belongs to negative perturbation, as it minimizes the training loss with the adversarial perturbations. Bootstrapping [130] is a typical robust loss based on label perturbation. It also belongs to negative perturbation as its perturbation is $\triangle y = \lambda(p - y)$, where p is the prediction output by the current trained model.

Some methods increase the losses of some samples and decrease those of others simultaneously. For instance, the losses of noisy-label training samples may be reduced, while those of clean samples may be increased in label smoothing.

3) Perturbation granularity: According to perturbation granularity, existing methods can be divided into sample-wise, class-wise, and corpus-wise.

- Sample-wise perturbation. In this division, each training sample has its own perturbation and different samples usually have distinct perturbations. The aforementioned Bootstrapping, and adversarial perturbation all belong to this division. The random cropping and masking also belong to this division.
- Class-wise perturbation. In this division, all the training samples in a category share the same perturbation, and different categories usually have distinct perturbations. Benz et al. [131] proposed a class-wise adversarial perturbation method. Wang et al. [132] introduced class-wise logit per-

turbation for semantic segmentation. Label smoothing also belongs to this division.

- Corpus-wise perturbation. In this division, all the training samples in the training corpus share only one perturbation. Shafahi et al. [133] pursued the universal adversarial perturbation for all the training samples, which has proven to be effective in various applications [134].

4) Assignment manner: The perturbation variables should be assigned before or during training. As presented in Fig. 7, there are four typical assignment manners to determine the perturbations.

- Rule-based assignment. In this manner, the perturbation is assigned according to pre-fixed rules. These rules are usually based on prior knowledge or statistical inspirations. In both label smoothing and Booststrapping loss, the label perturbation is determined according to manually defined formulas. In text classification, word replacement and random masking also obey manually developed rules.
- Regularization-based assignment. In this manner, a regularizer for the perturbation is added in the total loss. Take the logit perturbation as an example. A loss with regularization for logit perturbation can be defined as follows:

$$\mathcal{L} = \sum_i l(\mathcal{S}(v_i + \triangle v_i), y_i) + \lambda Reg(\triangle v_i), \tag{4}$$

  where $\mathcal{S}$ is the Softmax function, v_i is the logit vector for x_i, $\triangle v_i$ is the perturbation vector for v_i, and $Reg(\cdot)$ is the regularizer. Zhou et al. [135] introduced a novel adversarial perturbation way by leveraging smoothing regularization on adversarial perturbations. Wei et al. [136] proposed a sparse-regularized perturbation for video analysis.
- Loss-optimization-based assignment. This division is similar to the loss-optimization-based augmentation introduced in Section VI-B2. A new loss containing the perturbations is defined and the they are pursued by optimizing the loss. In the optimization procedure, only the perturbations are the variables to be optimized.
- Learning-based assignment. In this manner, the perturbation is assigned by leveraging a learning method. Three learning paradigms are usually applied as follows.
  - Self-supervised learning. This paradigm leverages self-supervised learning methodologies such as contrastive learning [137] to pursue the perturbations. Naseer et al. [138] constructed a self-supervised perturbation framework to optimize the feature distortion for an image.
  - Meta learning. This paradigm leverages meta-learning methodologies to pursue the perturbations using an additional meta dataset. It assumes that the perturbation $\triangle x$ (or $\triangle y$) for a training sample x (or its label y) is determined by the representation of x or factors such as training dynamics for x, which is described by $\triangle x = g(x, \eta(x))$, where $g(\cdot)$ can be a black-box neural network such as MLP; $\eta(x)$ represents the training dynamics for x. Li et al. [139] applied meta learning to directly optimize the covariant matrix used in ISDA, which is used to calculate the logit perturbation.
  - Reinforcement learning. This paradigm leverages reinforcement learning to pursue the perturbations without relying on additional data. Giovanni et al. [140] leveraged deep reinforcement learning to automatically generate realistic attack samples that can evade detection and train producing hardened models. Lin et al. [141] formulated the perturbation generation as a Markov decision process and optimized it by reinforcement learning.

Given a learning task, it is difficulty to directly judge which assignment manner is the most appropriate without a thorough and comprehensive understanding for the task. Each assignment manner has its own merits and defects.

D. Data weighting

Data weighting assigns a weight for each sample in loss calculation. It is among the most popular data optimization techniques in many scenarios, including fraud detection, portfolio selection, medical diagnosis, and fairness-aware learning. Three dividing dimensions are considered, namely, granularity, dependent factor, and assignment manner for weights.

1) Weighting granularity: According to the granularity of weights, existing weighting methods can be divided into sample-wise and category-wise. Noisy-label learning mainly adopts sample-wise methods [142], while imbalanced learning mainly adopts category-wise [143]. Weighting is also widely used in standard scenario, which is usually sample-wise.

2) Dependent factor: Dependent factor in this study denotes the factors that are leveraged to calculate the sample weights. Similar with the resampling introduced in Section VI-A, three factor types are usually considered, namely, category proportion, importance, and learning difficulty. As these concepts are introduced in Section VI-A and quite similar procedures are adopted, these factors are not detailed in this part. There are an increasing number of studies employing learning difficulty-based weighting. They can be further summarized according to which samples are learned first.

As samples with larger weights than others can be considered as having priority in training, learning difficulty-based weighting contains three basic folds, namely, easy-first, hard-first, and complicated.

- Easy-first. Easy samples are given higher weights than hard ones in this fold. There are two classical easy-first learning paradigms: curriculum learning [66] and self-paced learning [83]. These two paradigms assign larger weights to easy samples during the early training stage and gradually increase the weights of hard samples.
- Hard-first. Hard samples have higher weights than easy ones in this fold. The classical Focal loss is a typical hard-first strategy [67]. Zhang et al. [144] also assigned large weights on hard training samples.
- Complicated. In some weighting methods, the easy-first or the hard-first is combined with other weighting inspirations. In Balanced CL [145], on the basis of the easy-first mode, the selection of samples has to be balanced under certain constraints to ensure diversity.

Besides the three general ways, Zhou et al. [146] revealed some other priority types, including both-ends-first and varied manners during training. There also other dependent factors such as misclassified cost and those reflecting other concerns such as fairness and confidence [147], [148].

3) Assignment manner: Generally, there are four manners to assign weights for training samples as shown in Fig. 7.

- Rule-based assignment. This manner determines the sample weights according to theoretical or heuristic rules. For example, many methods assume that the category proportion is the prior probability. Consequently, the inverse of the category proportion is used as the weight based on the Bayesian rule. Cui et al. [143] established a theoretical framework for weight calculation based on the effective number theory in computation geometry. The classical Focal loss [67] heuristic defines the weight using $w = (1-p)^{\gamma}$, where p is the prediction on the ground-truth label and γ is a hyper-parameter. Han et al. [149] defined an uncertainty-based weighting manner for mixup. Importance weighting [150] is placed in this division.
- Regularization-based assignment. This method defines a new loss function which contains a weighted loss and a regularizer $Reg(\cdot)$ on the weights as follows:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} w_i l(f(x_i), y_i) + \lambda Reg(W), \tag{5}$$

  where $W = \{w_1, \cdots, w_N\}^T$ is the vector of weights on the N training samples. The classical self-paced learning, which mimics the mechanism of human learning from easy to hard gradually, is actually the regularization method defined as $Reg(W) = -|W|_1$ ($w_i \in \{0, 1\}$) [83].
- Adversarial optimization-based assignment. This manner pursues the sample weights by optimizing a defined objective function, which is similar to the pursing of the adversarial perturbation. For instance, Gu et al. [151] adversarially learned the weights of source domain samples to align the source and target domain distributions by maximizing the Wasserstein distance. Yi et al. [152] defined a maximal expected loss and obtained a simple and interpretable closed-form solution for samples' weights: larger weights are given to augmented samples with large losses.
- Learning-based assignment. Similar with that in data perturbation, learning-based assignment also usually applies meta learning or reinforcement learning to infer the sample weights. Ren et al. [153] firstly introduced meta learning for sample weighting in imbalanced learning and noisy-label learning. Shu et al. [154] utilized an MLP network to model the relationship between samples' characters and their weights, and then trained the network using meta learning. Li et al. [155] proposed meta learning-based weighting for pseudo-labeled samples in unsupervised domain adaptation. Ge et al. [156] used reinforcement learning to generate sample weights and combined the weights to train a recommendation system.

Weights assignment can be divided into static and dynamic methods. There are a few methods adopting static weighting [143], whereas most methods adopt dynamic weighting.

E. Data pruning

Data pruning is contrary to data augmentation. In this study, it is divided into dataset distillation and subset selection.

1) Dataset distillation: Dataset distillation is firstly proposed by Wang et al. [157] and it aims to synthesize a small typical training set from substantial data [158]. The synthesized dataset replaces the given dataset for efficient and accurate data-usage for the learning task. Following the division established by Sachdeva and McAuley [159], existing data distillation methods can be placed in four folds.

- Meta-model matching-based strategy. This strategy is firstly proposed by Wang et al. [157]. It performs an inner-loop optimization for a temporal optimal model based on the synthesized set and an outer-loop optimization for a temporal subset (i.e., the synthesized set) by turns. Loo et al. [160] utilized the light-weight empirical neural network Gaussian process kernel for the inner-loop optimization.
- Gradient matching-based strategy. This strategy [161] does not require to perform the inner-loop optimization as used in the meta-model matching-based strategy. Therefore, it is more efficient than the meta-model matching-based strategy. Kim et al. [162] further utilized spatial redundancy removing to accelerate the optimization process and gradients matching on the original dataset.
- Trajectory matching-based strategy. This strategy performs distillation by matching the training trajectories of models trained on the original and the pursued datasets [163]. Cui et al. [164] proposed a memory-efficient method which is avaliable for large datasets.
- Distribution matching-based strategy. This strategy performs the distillation by directly matching the distribution of the original dataset and the pursued dataset. Wang et al. [165] constructed a bilevel optimization strategy to jointly optimize a single encoder and summarize data.

There are some solutions which take alternative technical strategies, such as reinforcement learning [166], to solve the bi-level optimization in data distillation.

2) Subset selection: Subset selection aims to select the most useful samples from the original training set [33]. It does not generate any new samples. In Fig. 7, there are two divisions, including greedy search-based and mathematics-based.

In the greedy search-based strategy, the utility of each training sample is measured, and the subset is searched based on the utility rankings. According to the employed measures, existing methods can be divided into four categories, including difficulty-based, influence-based, value-based, and confidence-based. Meding et al. [167] utilized the misclassified rate by multiple classifiers as the learning difficulty to select samples. Feldman and Zhang [168] defined an influence score and a memorization score to measure each training sample. Birodkar et al. [6] employed clustering to select most valuable samples which are close to the cluster centers. Northcutt et al. [169] leveraged the confidence score to prune training samples.

Different from the greedy search strategy, some methods seek a global optimal subset according to a mathematical approach. Yang et al. [170] proposed a scalable framework to extract multiple mini-batch coresets from larger random subsets of training data by solving a submodular cover problem. Mirzasoleiman et al. [171] defined a monotonic function for coreset selection and proposed a generic algorithm with approximately linear complexity.

F. Other typical techniques

This study lists two representative techniques, including pure mathematical optimization and the combination of more than one aforementioned methods described in Section VI.

1) Pure mathematical optimization: This division refers to the manners that perform data optimization via pure mathematical optimization in the above-mentioned divisions.

The first typical scenario for pure mathematical optimization is the construction of a small-size yet high-quality dataset from the original training set. The tasks involving batch construction, meta data compiling in meta learning, or dataset distillation usually adopt mathematical optimization. Liu et al. [172] constructed a set variance diversity-based objective function for data augmentation and pursued the selection for a set of augmented samples via the maximization of the objective function in batch construction. Su et al. [72] established an objective function for meta data compiling and minimized it via submodular optimization. As introduced in Section VI-E, data pruning is usually performed based on pure mathematical optimization.

The second typical scenario is the regularized sample weighting or perturbation. The details are described in Sections VI-C4 and VI-D3. For instance, Li et al. [173] devised a new objective function for the label perturbation strength, which can also reduce the Bayes error rate during training. Meister et al. [174] constructed a general form of regularization that can derive a series of label perturbation methods.

The third typical scenario is the constrained optimization, which embeds prior knowledge into the constraints for weighting, perturbation, or pruning. For instance, Chai et al. [175] defined an optimization objective function with the constraints that each demographic group should have equal total weights in fairness-aware learning. The adversarial perturbation of multi-label learning is usually attained by solving constrained optimization problems [176].

2) Technique combination: Indeed, many learning algorithms do not employ a single data optimization technique. Instead, they combine different data optimization techniques. The following lists a few combination examples.

In data augmentation, many methods choose to generate samples in the first step and resample or reweight the samples in the second step. For instance, Cao et al. [177] dealt with grammatical error correction by using data weighting to balance the importance of each kind of augmented samples.

In data perturbation, different directions/granularity levels are usually combined in the same method. For example, adversarial perturbation belongs to the positive direction, while anti-adversarial perturbation belongs to the negative one. Zhao et al. [178] used both category-wise and sample-wise factors to infer the logit perturbation in imbalanced learning. Zhou et al. [49] combined adversarial and anti-adversarial perturbations and theoretically revealed the superiority of such strategy.

In data weighting, numerous methods combine it with data augmentation. Han et al. [149] combined uncertainty-based weighting and the augmentation method mixup. Chen et al. [119] combined effective number-based weighting and logit perturbation for imbalanced learning. Some other methods [67] combine different granularity levels or priority models.

G. Advantages and disadvantages of different techniques

This subsection attempts to provide a brief qualitative discussion of the advantages and disadvantages of our summarized optimization techniques.

Both data resampling and weighting have relatively low computational complexity. Different augmentation methods exhibit distinct computational complexities. For example, traditional image augmentation operations such as rotation and cropping have low complexity. However, model-based and loss-optimization-based explicit image augmentation methods generally have high computational complexity. The same applies to the perturbation technique. Methods such as label perturbation (e.g., label smoothing) are quite simple, whereas meta/reinforcement learning-based perturbation methods are typically complex. Data pruning has relatively high computational complexity, as most methods require solving an optimization problem.

Both data augmentation and perturbation can generate new data, whereas resampling and weighting cannot. Recent studies demonstrate that a moderate amount of synthetic data can help with model training [179]. Therefore, augmentation and perturbation seem to have more potential in the era of large models than resampling and weighting. However, not all augmented data are beneficial for large model training [180]. Thus, resampling and weighting have their own merits. Compared with augmentation, perturbation generally has lower computational complexity on average, as most perturbation methods do not actually generate new data. Of the two main technical strategies in pruning, dataset distillation is better for privacy preservation than subset selection [158]. Hu et al. [181] indicated that distillation outperforms subset selection in their empirical comparison.

The optimization goals and application scenarios of the five typical techniques differ. For example, data pruning aims to achieve the goal of 'efficiency', whereas data augmentation diverges from this goal. Most methods aim to improve model generalization, while adversarial perturbation primarily focuses on robustness, though it can enhance generalization in some image classification tasks.

It is important to note that the above comparative conclusion is not universally applicable to all learning tasks. There is no theoretical, universally established comparative conclusion for nearly any technique pair listed in Fig. 7 in the current literature.

VII. Data optimization theories

There are a large amount of studies focusing on the theoretical aspects of data optimization. It is quite challenging to arrange existing theoretical studies into a clear roadmap. This study summarizes existing studies in the following two dimensions, including formalization and explanation.

A. Formalization

In order to theoretically analyze and understand the data optimization methods, it is essential to establish mathematical formulations. Statistical modeling is the primary tool for their formalization [182]. Basic assumptions are usually relied on.

The most widely used assumptions for the statistical modeling include the following.

- Independent and identically distributed (I.I.D.)/Non-I.I.D. assumption. Most studies assume that each training sample is I.I.D. However, some studies focus on non-I.I.D. data. For example, Zheng et al. [183] theoretically investigated generative data augmentation in the non-I.I.D. setting.
- Gaussian distribution assumption. Many studies assume that data in each category conforms to a Gaussian distribution, which simplifies inference compared to other complicated distributions [111].
- Equal class CPD assumption. In many learning studies [184] excepting those for distribution drift, the class-conditional probability densities (CPD) of the training and testing sets are assumed to be identical.
- Uniform distribution assumption. In many studies, the distribution over categories in the testing set is assumed to be uniform. Some studies implicitly use this assumption by using modified losses such as the balanced accuracy or balanced test error [185], even if the category proportions in the test corpus are not identical.
- Linear boundary assumption. In many studies [186], [187], the decision boundary of the involved classifier is assumed to be linear. The decision boundary between two categories under the cross-entropy loss is linear.

Based on these assumptions, the data optimization problems are usually formalized into probabilistic, constrained optimization, or regularization-based problems. For example, Xu et al. [188] investigated importance weighting for covariate-shift generalization based on probabilistic analysis. Chen et al. [189] defined the classification accuracy based on posterior probability for zero-shot learning. Qraitem et al. [190] formalized a constrained linear program problem to investigate the effect of data resampling. Roh et al. [191] formulated a combinatorial optimization problem for the unbiased selection of samples in the presence of data corruption. In classical weighting paradigm such as SPL, data weighting is directly formalized in the optimization object consisting of the weighted loss and a regularizer. Zhang et al. [192] defined a re-weighted score function consisting of weighted loss and a sparsity regularization for causal discovery.

Jiang et al. [193] proposed a new adversarial perturbation generation method by adding a diversity-based regularization which measures the diversity of candidates. Hounie et al. [194] proposed a constrained learning problem for automatic data augmentation by combining conventional training loss and risk constraints.

B. Explanation

Most theoretical studies on data optimization aim to explain why the existing methods are effective or ineffective.

In data perception, researchers usually conducted theoretical analysis on the role of one typical data measure or leveraged the measure to understand the training process of DNNs. Doan et al. [195] conducted a theoretical analysis of catastrophic forgetting in continuous learning. Chatterjee et al. [196] utilized the perception on gradients to explain the generalization of deep learning.

In data resampling, existing theoretical studies focus on importance sampling for deep learning. Katharopoulos and Fleuret [93] derived an estimator of the variance reduction achieved with importance sampling in deep learning. Wang et al. [197] proposed an unweighted data sub-sampling method, and proved that the subset-model acquired through the method outperforms the full-set-model.

In data augmentation, more and more theoretical studies are performed. Dao et al. [198] established a theoretical framework for data augmentation and revealed that data augmentation can be approximated by first-order feature averaging and second-order variance regularization. Wu and He [199] investigated the theoretical issues for adversarial perturbations for multi-source domain adaptation.

In data perturbation, most theoretical studies focus on the adversarial perturbation. Yi et al. [200] investigated the models trained by adversarial training on OOD data. Some studies delved into the theoretical justification for label and logit perturbation. Li et al. [125] theoretically analyzed the usefulness of logit adjustment in dealing with the class imbalance.

In data weighting, Byrd and Lipton [91] investigated the role of importance in deep learning. Fang et al. [201] discussed the limitations of importance weighting and found that it suffers from a circular dependency. Weinshall et al. [202] proved that the convergence rate of an ideal curriculum learning method is monotonically increasing with the samples' learning difficulty.

In data pruning, theoretical studies are relatively limited. Zhu et al. [203] revealed that distilled data lead to networks that are not calibratable. Dong et al. [204] theoretically revealed the connection between dataset distillation and differential privacy.

There are also studies which aim to reveal the intrinsic connections between two different technical paths. For instance, regularization is a widely used technique in deep learning [205], and several typical data optimization techniques are revealed to be a regularization method [42], [206], [207]. Therefore, intrinsic connections among these techniques can be established, which enlightens a better understanding of the involved technical paths and can envision novel methods.

VIII. Connections among different techniques

Five aspects, namely, perception, application scenarios, similarity/opposition, theories, and data types, connect different methods within a technical path or across different paths.

A. Connections via data perception

Data perception is the first (explicit or implicit) step in the data optimization pipeline. Methods along different technical paths introduced in Sections VI-A-VI-F may choose the same or similar quantities in perception. Therefore, quantities for data perception connect different methods. For example, many data optimization methods are on the basis of training loss in resampling [95], augmentation [49], perturbation [123], weighting [83], and subset selection [167]. Gradient is widely used in resampling [93], augmentation [224], perturbation [49], weighting [59], and dataset distillation [161]. Other quantities including margin and uncertainty are also used in different optimization techniques.

TABLE I
SOME DATA OPTIMIZATION METHODS IN NOISY-LABEL LEARNING.

Datasets	Resampling	Augmentation	Perturbation	Weighting	Dataset pruning
CIFAR10	[95], [197], [208]	[209], [210], [211], [212]	[126], [130]	[83], [213], [214], [215]	[216], [217], [218]
CIFAR100	[95], [197], [208]	[209], [210], [211], [212]	[126], [130]	[83], [214], [215]	[217], [218]
Clothing1M	[197]	[209], [210], [212]	[126], [130]	[83], [215]	[218]
WebVision	[53]	[212], [210]	[64]	[213], [214], [215]	[217]

TABLE II
SOME DATA OPTIMIZATION METHODS IN IMBALANCED LEARNING.

Datasets	Resampling	Augmentation	Perturbation	Weighting	Dataset pruning
CIFAR10-LT	[219]	[219], [220]	[124], [139]	[67], [143]	[72], [221]
CIFAR100-LT	[222]	[119], [220]	[124], [139]	[67], [143]	[72], [221]
iNaturallist	[222]	[119], [220]	[124], [139]	[67], [143]	[72]
ImageNet-LT	[222]	[119], [220]	[124], [139]	[67], [143]	[72], [223]

The utilization of the same or similar perception quantities demonstrates that these methods have the same or similar heuristic observations or theoretical inspirations.

B. Connections via application scenarios

Most data optimization methods can be leveraged for the application scenarios discussed in Section IV-B.

One of the most focused scenarios of data optimization methods is noisy-label learning. Many classical methods are from resampling, augmentation, weighting, or perturbation. These are also dataset distillation studies for noisy-label datasets [225]. Table I shows some representative data optimization methods for noisy-label learning on five benchmark datasets CIFAR10 [226], CIFAR100 [226], Clothing1M [227], and WebVision [228].

Imbalanced learning is also among the most focused scenarios. Nearly all the listed data optimization techniques have been used in imbalanced learning. Table II shows some representative methods for imbalanced learning on four benchmark datasets CIFAR10-LT [143], CIFAR100-LT [143], iNaturalist [229], and ImageNet-LT [230]. There are some studies employing more than one type of data optimization techniques such as ReMix [219], which combines resampling and augmentation, in Table II.

Robust learning for adversarial attacks is another typical scenario. Karimireddy and Jaggi [231] employed resampling to design robust models. Data weighting [144] and dataset distillation [232] are also used in robust learning.

C. Connections via similarity/opposition

The similar and opposite relationships existing among the five technical paths are introduced in Section VI.

Data resampling and weighting are closely related techniques, as their key steps are nearly the same. Therefore, in many studies on noisy-label learning and imbalanced learning, these two techniques are often considered as a single strategy.

Although data pruning and augmentation are opposite to each other, they have consistent ultimate goals in learning tasks. They are overlapped in terms of employed methodologies as shown in Fig. 8. It is believable that more intrinsic connections can be explored for them.

In the data resampling, weighting, and perturbation, the assignment manners for the sampling rate, weighting score, and perturbation variable are quite similar. In addition to the classical importance score, both meta learning [233] and adversarial strategy [100] have also been used in data resampling. Regularization-based manner is used in nearly all the data optimization paths except resampling. Due to space constraints, methods with different assignment manners are not summarized in a table as those in Section VIII-B.

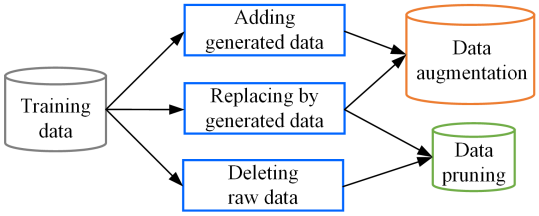

Fig. 8. Connection between augmentation and pruning.

There are other opposite relationships, such as undersampling vs. oversampling, easy-first weighting vs. hard-first weighting, positive perturbation vs. negative perturbation, and explicit augmentation vs. implicit augmentation. Both methodologies in these opposite relationships have been demonstrated to be effective in previous literature, aligning with the proverb "All roads lead to Rome".

D. Connections via theory

There are some common theoretical issues, analyses, and conclusions heavily influencing most data optimization techniques. They are the natural connections among different techniques. Several examples are listed as follows:

- Theoretical issues in data perception. A solid theoretical basis for data perception in data optimization is lacking, even though most data optimization methods implicitly or explicitly rely on the perception for the training data. For instance, many methods from resampling, weighting, and perturbation are based on dividing samples into easy and hard. Nevertheless, there is not yet a widely accepted learning difficulty measure in the literature.
- Probabilistic density (ratio) estimation. Many data optimization methods, especially resampling and weighting, heavily rely on the probabilistic density (ratio) estimation. The most representative method is the importance sampling. In learning difficulty-based weighting, the probabilistic density ratio, in terms of learning difficulty, is revealed to determine the priority mode [146], namely, easy/medium/hard-first.
- Regularization-based explanation. Many data optimization methods are considered as a type of regularization, including data augmentation and perturbation. In these methods, data

optimization performs implicit model regularization other than explicit regularization that directly works on model parameters. Regularization is not always beneficial as over-regularization may occur. Li et al. [234] pointed out that large amount of augmented noisy data could lead to over-regularization and proposed a new augmentation method.

- Generalization bound for data optimization. Many studies choose to deduce a mathematical bound in terms of the variables related to data optimization. This manner can theoretically explain the utility of data optimization. Xiao et al. [235] derived stability-based generalization bounds for SGD on the loss with adversarial perturbations. Xu et al. [187] established a new generalization bound that reflects how importance weighting leads to the interplay between the empirical risk and the distribution deviation.

The progress in each of the above theoretical aspects will promote the advancement of many data optimization methods in different technical paths.

E. Connections via data types

The data types of the datasets listed in Tables I and II are images. Other common data types include text, time series, graphs, and tabular data. Theoretically, most existing data optimization techniques can be applied to these data types with adaptations based on their specific characteristics. Researchers from various fields, including computer vision, natural language processing, graph neural networks, and time series analysis, have contributed to data optimization methodologies. Generally, resampling methods for images, text, time series, and tabular data are nearly the same, as they are independent of the specific data types. Resampling on graphs should adopt special strategies, as nodes are interconnected. Data augmentation methods for different data types usually vary significantly in their implementation details, as they are heavily dependent on the specific data types. The differences among data perturbation methods for different data types depend on the perturbation targets. Label, logit, feature, and gradient perturbation strategies for different data types are usually identical or quite similar, whereas sample perturbation strategies for different data types may vary significantly. Weighting methods across different data types vary slightly. Dataset distillation-based pruning for different data types also varies significantly, as it requires generating new samples.

Although many methods in an optimization technique can be applied indiscriminately to different data types, each data type has specific preferences for certain methods. For example, sample augmentation with adversarial learning achieves good performance on images, whereas it may fail on text.

IX. Future directions

A. Principles of data optimization

Up till now, there has been no consensus theoretical framework that is suitable for all or most technical paths. There are some studies aiming to establish the connection between two different technical paths, such as resampling vs. weighting [236]. Many open problems or controversies remain unsolved. For example, there is no ideal answer for which resampling strategy should be employed first: oversampling or undersampling? Megahed et al. [88] suggested that undersampling should be used firstly, whereas Xie et al. [111] demonstrated that oversampling is effective. Likewise, although Zhou et al. [146] provided an initial answer for the choice of easy-first and hard-first weighting strategies, a solid theoretical framework is still lacking in their study.

Moreover, even for a single data optimization method, multiple explanations from different views may exist. The explanation for label smoothing is a typical example. At least four studies provide empirical or theoretical explanations for it [174], [237]–[239]. Regarding the effectiveness of adversarial samples, some researchers have pointed out that adversarial samples are useful features [240], while some other researchers investigated it in terms of gradient regularization [241].

Consequently, the construction of the data optimization principles is of great importance, as it can promote the establishing of a unified and solid theoretical framework which can be used to analyze and understand of each data optimization technical. There have been studies on the first principle for the design of DNNs [242]. To explore the principles for data optimization, a unified mathematical formalization is required and large-scale empirical studies (e.g., [243]) will be helpful.

B. Interpretable data optimization

Interpretable data optimization refers to the explanation for the involved data optimization techniques in terms of how and which aspects they affect the training process of DNNs. Although interpretable deep learning receives much attention in recent years [244], it focuses on DNN models other than the training processing in which data optimization techniques are involved. Interpretable data optimization is an under-explored research topic and there are limited studies on this topic [245], [246]. The well explanation of how and which aspects of a data optimization method affects a specific training process is significant beneficial for the design or selecting of more effective optimization methods.

C. Human-in-the-loop data optimization

Recently, human-in-the-loop (HITL) deep learning receives increasing attention in the AI community [247]. With out human's participants, high-quality samples are not intractable to obtain. Naturally, HITL data optimization can also be beneficial for deep learning. Collins et al. [248] investigated HITL mixup and indicated that collating humans' perceptions on augmented samples could impact model performance. Wallace et al. [249] proposed HITL adversarial generation, where human authors are guided to break models. Overall, research on HITL data optimization is in the early stage.

D. Data optimization for new challenges

New challenges are constantly emerging in deep learning applications. We take the following challenges as examples to illustrate the future direction of data optimization:

- Open-world learning. This learning scenario confronts the challenge of out-of-distribution (OOD) samples. Wu et al. [250] investigated the learning issue when both OOD and noisy samples exist. Some other studies investigate ODD under imbalanced learning [251].

- Large-model training. Large models especially the large language models have achieved great success in recent years. Data optimization can also take effect in large models' training. Wei et al. [252] investigated the condensation of prompts and promising results are obtained.
- Multi-modal learning. Multi-modal data are avaliable in more and more real tasks [253]. Consequently, many learning tasks are actually multi-modal learning. As each sample consists of raw data/features from different modalities, the data perception for multi-modal samples should be different from that for single-modal samples. The data optimization methods are likewise different from conventional methods.

E. Data optimization for AI security

With the increasing impact of AI technology on society, AI security, such as adversarial robustness, model trustworthiness, and data privacy, is becoming increasingly important. Data optimization has also been used to address AI security issues. Adversarial perturbation has proven to be an effective technique for improving the adversarial robustness of DNN models. It can also be used to implant backdoors into AI systems [254]. Data is crucial for trustworthy AI. Liang et al. [255] pointed out that data critically affects the trustworthiness of a model, while the design and sculpting of data used to develop AI often rely on bespoke manual work. In terms of data privacy and copyright issues, dataset distillation is a promising technique, as it generates new training data [256]. Yu et al. [257] revealed that the privacy risk of models trained with data augmentation could be largely underestimated. Li et al. [258] evaluated several data augmentation methods in terms of privacy attack and suggested that some methods are effective in reducing the vulnerability to such privacy attacks.

As AI security is a major concern for AI applications, data optimization for AI security will present additional challenges and create research opportunities for the entire community.

X. Conclusions

This paper aims to summarize a wide range of learning methods within an independent deep learning realm, namely, data optimization. A taxonomy for data optimization, as well as fine-granularity sub-taxonomies, is established for existing studies on data optimization. Connections among different methods are discussed, and potential future directions are presented. It is noteworthy that many classical methods, such as dropout, are essentially data optimization methods. In our future work, we will explore a more fundamental and unified viewpoint on data optimization, and develop a more comprehensive taxonomy to incorporate more classical methods.

References

[1] S. E. Whang *et al.*, "Data collection and quality challenges in deep learning: A data-centric ai perspective," *The VLDB Journal*, vol. 32, no. 4, pp. 791–813, 2023.

[2] M. H. Jarrahi *et al.*, "The principles of data-centric ai," *Communications of the ACM*, vol. 66, no. 8, pp. 84–92, 2023.

[3] Y. Liang *et al.*, "Multi-view graph learning by joint modeling of consistency and inconsistency," *IEEE TNNLS*, pp. 1–15, 2022.

[4] P. Zhu *et al.*, "Latent heterogeneous graph network for incomplete multi-view learning," *IEEE TMM*, vol. 25, pp. 3033–3045, 2023.

[5] L. Brigato *et al.*, "A close look at deep learning with small data," in *ICPR*, 2021, pp. 2490–2497.

[6] V. Birodkar *et al.*, "Semantic redundancies in image-classification datasets: The 10% you don't need," *arXiv:1901.11409*, 2019.

[7] Y. Yu *et al.*, "Can data diversity enhance learning generalization?" in *COLING*, 2022, pp. 4933–4945.

[8] J. Lu *et al.*, "Learning under concept drift: A review," *IEEE TKDE*, vol. 31, no. 12, pp. 2346–2363, 2018.

[9] W. Wang *et al.*, "Towards a robust deep neural network against adversarial texts: A survey," *IEEE TKDE*, vol. 35, no. 3, pp. 3159–3179, 2023.

[10] Y. Wu *et al.*, "On convexity and bounds of fairness-aware classification," in *WWW*, 2019, pp. 3356–3362.

[11] P. Xiong *et al.*, "Towards a robust and trustworthy machine learning system development: An engineering perspective," *JISA*, vol. 65, p. 103121, 2022.

[12] H. Zhang *et al.*, "mixup: Beyond empirical risk minimization," *ICLR*, 2018.

[13] R. Yao *et al.*, "Compensation learning," *arXiv:2107.11921*, 2022.

[14] X. Wang *et al.*, "Deep generative mixture model for robust imbalance classification," *IEEE TPAMI*, vol. 45, no. 3, pp. 2897–2912, 2023.

[15] H. He *et al.*, "Learning from imbalanced data," *IEEE TKDE*, vol. 21, no. 9, pp. 1263–1284, 2009.

[16] Y. Zhang *et al.*, "Deep long-tailed learning: A survey," *IEEE TPAMI*, vol. 45, no. 9, pp. 10 795–10 816, 2023.

[17] G. Algan *et al.*, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowledge-Based Systems*, vol. 215, no. 5, p. 106771, 2021.

[18] H. Song *et al.*, "Learning from noisy labels with deep neural networks: A survey," *IEEE TNNLS*, pp. 1–19, 2022.

[19] X. Cao *et al.*, "A survey of learning on small data," *arXiv:2207.14443*, 2022.

[20] Y. Wang *et al.*, "Generalizing from a few examples: A survey on few-shot learning," *ACM CSUR*, vol. 53, no. 3, pp. 1–34, 2020.

[21] L. Yuan *et al.*, "Recent advances in concept drift adaptation methods for deep learning," in *IJCAI*, 2022, pp. 5654–5661.

[22] A. Diez-Olivan *et al.*, "Adaptive dendritic cell-deep learning approach for industrial prognosis under changing conditions," *IEEE TII*, vol. 17, no. 11, pp. 7760–7770, 2021.

[23] S. H. Silva *et al.*, "Opportunities and challenges in deep learning adversarial robustness: A survey," *arXiv:2007.00753*, 2020.

[24] J. Xu *et al.*, "Robustness of deep learning models on graphs: A survey," *AI Open*, vol. 2, pp. 69–78, 2021.

[25] S. Goyal *et al.*, "A survey of adversarial defenses and robustness in nlp," *ACM CSUR*, vol. 55, no. 14s, pp. 1–39, 2023.

[26] N. Mehrabi *et al.*, "A survey on bias and fairness in machine learning," *ACM CSUR*, vol. 54, no. 6, pp. 1–35, 2021.

[27] A. Petrović *et al.*, "Fair: Fair adversarial instance re-weighting," *Neurocomputing*, vol. 476, pp. 14–37, 2022.

[28] S. K. Devitt, "Trustworthiness of autonomous systems," *Foundations of trusted autonomy*, pp. 161–184, 2018.

[29] D. Kaur *et al.*, "Trustworthy artificial intelligence: A review," *ACM CSUR*, vol. 55, no. 2, pp. 1–38, 2022.

[30] B. Wu *et al.*, "Trustworthy graph learning: Reliability, explainability, and privacy protection," *ACM KDD*, 2022.

[31] C. Fang *et al.*, "Combating noisy labels in long-tailed image classification," *arXiv:2209.00273*, 2022.

[32] M. Singh *et al.*, "An empirical study of accuracy, fairness, explainability, distributional robustness, and adversarial robustness," in *KDD Workshop*, 2021.

[33] Z. Wan *et al.*, "A survey of data optimization for problems in computer vision datasets," *arXiv:2210.11717*, 2022.

[34] D. Zha *et al.*, "Data-centric ai: Techniques and future perspectives," in *ACM KDD*, 2023, p. 5839–5840.

[35] S. Whang *et al.*, "Data collection and quality challenges in deep learning: a data-centric ai perspective," *The VLDB Journal*, vol. 32, pp. 791–813, 2023.

[36] D. Zha *et al.*, "Data-centric artificial intelligence: A survey," *arXiv:2303.10158v3*, 2023.

[37] G. Fu *et al.*, "Making waves: Towards data-centric water engineering," *Water Research*, vol. 256, p. 121585, 2024.

[38] I. Pan *et al.*, "Data-centric engineering: integrating simulation, machine learning and statistics. challenges and opportunities," *Chemical Engineering Science*, vol. 249, p. 117271, 2022.

[39] N. Chawla *et al.*, "Smote: synthetic minority over-sampling technique," *JAIR*, vol. 16, no. 1, pp. 321–357, 2002.

[40] Y. Luo *et al.*, "$\mathcal{G}$-softmax: Improving intraclass compactness and interclass separability of features," *IEEE TNNLS*, vol. 31, no. 2, pp. 685–699, 2020.

[41] A. Damian *et al.*, "Label noise sgd provably prefers flat global minimizers," in *NeurIPS*, 2021, pp. 27 449–27 461.

[42] Y. Wang *et al.*, "Implicit semantic data augmentation for deep networks," in *NeurIPS*, 2019, pp. 12 635–12 644.

[43] M. Wang *et al.*, "Meta balanced network for fair face recognition," *IEEE TPAMI*, 2021.

[44] S. Fujii *et al.*, "Data augmentation by selecting mixed classes considering distance between classes," *arXiv:2209.05122*, 2022.

[45] C.-Y. Chuang *et al.*, "Fair mixup: Fairness via interpolation," in *ICLR*, 2021.

[46] L. E. Celis *et al.*, "Fair classification with adversarial perturbations," in *NeurIPS*, 2021, pp. 8158–8171.

[47] B. Yan *et al.*, "Forml: Learning to reweight data for fairness," *arXiv:2202.01719*, 2022.

[48] S. Lee *et al.*, "Graddiv: Adversarial robustness of randomized neural networks via gradient diversity regularization," *TPAMI*, vol. 45, no. 2, pp. 2645–2651, 2023.

[49] X. Zhou *et al.*, "Combining adversaries with anti-adversaries in training," in *AAAI*, 2023.

[50] B. Liu *et al.*, "The devil is in the margin: Margin-based label smoothing for network calibration," in *CVPR*, 2022, pp. 80–88.

[51] J. Mukhoti *et al.*, "Calibrating deep neural networks using focal loss," in *NeurIPS*, 2020, pp. 15 288–15 299.

[52] M. P. Naeini *et al.*, "Obtaining well calibrated probabilities using bayesian binning," in *AAAI*, 2015.

[53] S. Kim *et al.*, "Coreset sampling from open-set for fine-grained self-supervised learning," in *CVPR*, 2023, pp. 7537–7547.

[54] L. Dunlap *et al.*, "Diversify your vision datasets with automatic diffusion-based augmentation," *arXiv:2305.16289*, 2023.

[55] Z. Ye *et al.*, "Infusing definiteness into randomness: Rethinking composition styles for deep image matting," *AAAI*, 2023.

[56] W. Yang *et al.*, "Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages," *arXiv:2305.18098*, 2023.

[57] X. Liu *et al.*, "Adversarial training for large neural language models," *arXiv:2004.08994*, 2020.

[58] J. Huang *et al.*, "O2u-net: A simple noisy label detection approach for deep neural networks," in *ICCV*, 2019, pp. 3326–3334.

[59] B. Li *et al.*, "Gradient harmonized single-stage detector," in *AAAI*, 2019, pp. 8577–8584.

[60] C.-B. Zhang *et al.*, "Delving deep into label smoothing," *IEEE TIP*, vol. 30, pp. 5984–5996, 2021.

[61] X. Ning *et al.*, "Hyper-sausage coverage function neuron model and learning algorithm for image classification," *Pattern Recognition*, vol. 136, p. 109216, 2023.

[62] J. Lin *et al.*, "Measuring the effect of training data on deep learning predictions via randomized experiments," in *ICML*, 2022, pp. 13 468–13 504.

[63] S. Shrivastava *et al.*, "Datasetequity: Are all samples created equal? in the quest for equity within datasets," in *ICCV*, 2023, pp. 4417–4426.

[64] A. Iscen *et al.*, "Learning with neighbor consistency for noisy labels," in *CVPR*, 2022, pp. 4672–4681.

[65] R. Wang *et al.*, "Tackling the imbalance for gnns," in *IJCNN*, 2022.

[66] Y. Bengio *et al.*, "Curriculum learning," in *ICML*, 2009, pp. 41–48.

[67] T. Lin *et al.*, "Focal loss for dense object detection," in *CVPR*, 2017, pp. 2999–3007.

[68] W. Zhu *et al.*, "Exploring the learning difficulty of data: Theory and measure," *arXiv:2205.07427*, 2022.

[69] M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information fusion*, vol. 76, pp. 243–297, 2021.

[70] A. Kendall *et al.*, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NeurIPS*, 2017, pp. 5574–5584.

[71] A. Kumar *et al.*, "Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation," in *NAACL*, 2019, pp. 3609–3619.

[72] F. Su *et al.*, "Submodular meta data compiling for meta optimization," *ECML/PKDD*, 2022.

[73] D. Friedman *et al.*, "The vendi score: A diversity evaluation metric for machine learning," *arXiv:2210.02410*, 2023.

[74] Y. Pang *et al.*, "Imbalanced ensemble learning leveraging a novel data-level diversity metric," *Pattern Recognition*, vol. 19, no. 110886, 2025.

[75] O. Wu, "Rethinking class imbalance in machine learning," *arXiv:2305.03900*, 2023.

[76] S. Swayamdipta *et al.*, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *EMNLP*, 2020.

[77] Z. Wang *et al.*, "A comprehensive survey of forgetting in deep learning beyond continual learning," *arXiv:2307.09218*, 2023.

[78] L. S. Shapley, "A value for n-person games," in *In Contributions to the Theory of Games*, 1953, pp. 307–317.

[79] A. Ghorbani *et al.*, "Data shapley: Equitable valuation of data for machine learning," in *ICML*, 2019, pp. 2242–2251.

[80] K. F. Jiang *et al.*, "Opendataval: a unified benchmark for data valuation," in *NeurIPS*, 2023.

[81] C. Dong *et al.*, "Data profiling for adversarial training: On the ruin of problematic data," *arXiv:2102.07437v1*, 2021.

[82] Z. Hammoudeh *et al.*, "Training data influence analysis and estimation: A survey," *arXiv:2212.04612*, 2023.

[83] M. P. Kumar *et al.*, "Self-paced learning for latent variable models," *NeurIPS*, pp. 1–9, 2010.

[84] M. Toneva *et al.*, "An empirical study of example forgetting during deep neural network learning," *ICLR*, 2019.

[85] E. Arazo *et al.*, "Unsupervised label noise modeling and loss correction," in *ICML*, 2019.

[86] C. Hu *et al.*, "Mild: Modeling the instance learning dynamics for learning with noisy labels," *arXiv:2306.11560*, 2023.

[87] Y. Li *et al.*, "Repair: Removing representation bias by dataset resampling," *CVPR*, 2019.

[88] F. M. Megahed *et al.*, "The class imbalance problem," *Nature Methods*, vol. 18, pp. 1270–1272, 2021.

[89] J. Cui *et al.*, "Reslt: Residual learning for long-tailed recognition," *IEEE TPAMI*, vol. 45, no. 3, pp. 3695–3706, 2023.

[90] I. L. . F. Hutter, "Online batch selection for faster training of neural networks," *ICLR Workshop*, 2016.

[91] J. Byrd *et al.*, "What is the effect of importance weighting in deep learning?" in *ICML*, 2019, pp. 872–881.

[92] Q. Liu *et al.*, "Black-box Importance Sampling," in *AISTATS*, 2017, pp. 952–961.

[93] A. Katharopoulos *et al.*, "Not all samples are created equal: Deep learning with importance sampling," in *ICML*, 2018, pp. 2525–2534.

[94] T. B. Johnson *et al.*, "Training deep models faster with robust, approximate importance sampling," in *NeurIPS*, 2018.

[95] X. J. Gui *et al.*, "Towards understanding deep learning from noisy labels with small-loss criterion," *IJCAI*, 2021.

[96] V. Nguyen *et al.*, "How to measure uncertainty in uncertainty sampling for active learning," *Machine Learning*, vol. 111, pp. 89–122, 2022.

[97] J. Mena *et al.*, "A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective," *ACM CSUR*, vol. 54, no. 9, pp. 1–35, 2021.

[98] A. Aljuhani *et al.*, "Uncertainty aware sampling framework of weak-label learning for histology image classification," in *MICCAI*, 2022, pp. 366–376.

[99] D. Ting *et al.*, "Optimal subsampling with influence functions," in *NeurIPS*, 2018, pp. 3650–3659.

[100] Y. Li *et al.*, "Background data resampling for outlier-aware classification," in *CVPR*, 2020, pp. 13 218–13 227.

[101] J. Zhang *et al.*, "Undersampling near decision boundary for imbalance problems," in *ICMLC*, 2019, pp. 1–8.

[102] M. Sun *et al.*, "Autosampling: Search for effective data sampling schedules," in *ICML*, 2017, p. 9923–9933.

[103] M. Bayer *et al.*, "A survey on data augmentation for text classification," *ACM CSUR*, vol. 55, no. 7, pp. 1–39, 2022.

[104] C. Shorten *et al.*, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 60, 2019.

[105] B. Li *et al.*, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, 2022.

[106] P. Li *et al.*, "A simple feature augmentation for domain generalization," in *ICCV*, 2021, pp. 8886–8895.

[107] P. Chu *et al.*, "Feature space augmentation for long-tailed data," in *ECCV*, 2020, pp. 694–710.

[108] H. Lee *et al.*, "Self-supervised label augmentation via input transformations," in *ICML*, 2020, pp. 5714–5724.

[109] F. Huang *et al.*, "Adversarial and isotropic gradient augmentation for image retrieval with text feedback," *IEEE TMM*, pp. 1–12, 2022.

[110] D. Dablain *et al.*, "Deepsmote: Fusing deep learning and smote for imbalanced data," *IEEE TNNLS*, vol. 34, no. 9, pp. 6390–6404, 2023.

[111] Y. Xie *et al.*, "Gaussian distribution based oversampling for imbalanced data classification," *IEEE TKDE*, vol. 32, no. 2, pp. 667–679, 2022.

[112] I. Goodfellow *et al.*, "Generative adversarial nets," in *NeurIPS*, 2014.

[113] Z. Yang *et al.*, "Ts-gan: Time-series gan for sensor-based health data augmentation," *ACM TOCH*, vol. 4, no. 2, pp. 1–21, 2022.

[114] L. Yang *et al.*, "Diffusion models: A comprehensive survey of methods and applications," *ACM CSUR*, 2023.

[115] P. McNamee *et al.*, "An extensive exploration of back-translation in 60 languages," in *ACL Findings*, 2023, pp. 8166–8183.

[116] M. Pagliardini *et al.*, "Improving generalization via uncertainty driven perturbations," *arXiv:2202.05737*, 2022.

[117] Z. Mai *et al.*, "Metamixup: Learning adaptive interpolation policy of mixup with metalearning," *IEEE TNNLS*, vol. 33, no. 7, pp. 3050–3064, 2021.

[118] T. Qin *et al.*, "Automatic data augmentation via deep reinforcement learning for effective kidney tumor segmentation," in *ICASSP*, 2020, pp. 1419–1423.

[119] X. Chen *et al.*, "Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification," in *AAAI*, Online, February 2022, pp. 356–364.

[120] X. Zhou *et al.*, "Implicit counterfactual data augmentation for deep neural networks," *arXiv:2304.13431*, 2023.

[121] A. Jeddi *et al.*, "Learn2perturb: An end-to-end feature perturbation learning to improve adversarial robustness," in *CVPR*, 2020.

[122] M. Shu *et al.*, "Encoding robustness to image style via adversarial feature perturbations," in *NeurIPS*, 2021, pp. 28 042–28 053.

[123] M. Li *et al.*, "Logit perturbation," in *AAAI*, 2022, pp. 10 388–10 396.

[124] A. K. Menon *et al.*, "Long-tail learning via logit adjustment," in *ICLR*, 2021.

[125] M. Li *et al.*, "Class-level logit perturbation," *IEEE TNNLS*, 2023.

[126] C. Szegedy *et al.*, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[127] J. Wang *et al.*, "Reinforcement learning with perturbed rewards," in *AAAI*, 2020, pp. 6202–6209.

[128] A. Orvieto *et al.*, "Anticorrelated noise injection for improved generalization," in *ICML*, 2022, pp. 17 094–17 116.

[129] D. Wu *et al.*, "Adversarial weight perturbation helps robust generalization," in *NeurIPS*, 2020, pp. 2958–2969.

[130] S. Reed *et al.*, "Training deep neural networks on noisy labels with bootstrapping," in *ICLR Workshop*, 2015.

[131] P. Benz *et al.*, "Universal adversarial training with class-wise perturbations," in *ICME*, 2021, pp. 1–6.

[132] Y. Wang *et al.*, "Balancing logit variation for long-tailed semantic segmentation," in *CVPR*, 2023, pp. 19 561–19 573.

[133] A. Shafahi *et al.*, "Universal adversarial training," in *CVPR*, 2017, pp. 5636–5643.

[134] A. Chaubey *et al.*, "Universal adversarial perturbations: A survey," *arXiv:2005.08087*, 2020.

[135] W. Zhou *et al.*, "Transferable adversarial perturbations," in *ECCV*, 2018, pp. 452–467.

[136] X. Wei *et al.*, "Sparse adversarial perturbations for videos," in *AAAI*, 2019, pp. 8973–8980.

[137] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[138] M. Naseer *et al.*, "A self-supervised approach for adversarial robustness," in *CVPR*, 2020, pp. 262–271.

[139] S. Li *et al.*, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," in *CVPR*, 2021, pp. 5212–5221.

[140] G. Apruzzese *et al.*, "Deep reinforcement adversarial learning against botnet evasion attacks," *IEEE TNSE*, vol. 17, no. 4, pp. 1975–1987, 2020.

[141] B. Lin *et al.*, "Adversarial reinforced instruction attacker for robust vision-language navigation," *IEEE TPAMI*, vol. 44, no. 10, pp. 7175–7189, 2022.

[142] T. Castells *et al.*, "Superloss: A generic loss for robust curriculum learning," in *NeurIPS*, 2020, pp. 1–12.

[143] Y. Cui *et al.*, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9260–9269.

[144] J. Zhang *et al.*, "Geometry-aware instance-reweighted adversarial training," in *ICLR*, 2021.

[145] P. Soviany, "Curriculum learning with diversity for supervised computer vision tasks," in *ICML Workshop*, 2020.

[146] X. Zhou *et al.*, "Which samples should be learned first: Easy or hard?" *IEEE TNNLS*, pp. 1–15, 2023.

[147] W. Zhang *et al.*, "Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition," in *CVPR*, June 2019, pp. 7373–7382.

[148] C. Northcutt *et al.*, "Confident learning: Estimating uncertainty in dataset labels," *JAIR*, vol. 70, pp. 1373–1411, 2021.

[149] Z. Han *et al.*, "Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup," in *NeurIPS*, 2022, pp. 37 704–37 718.

[150] T. Liu *et al.*, "Classification with noisy labels by importance reweighting," *IEEE TPAMI*, vol. 38, no. 3, p. 447–461, 2023.

[151] X. Gu *et al.*, "Adversarial reweighting for partial domain adaptation," in *NeurIPS*, 2021, pp. 14 860–14 872.

[152] M. Yi *et al.*, "Reweighting augmented samples by minimizing the maximal expected loss," in *ICLR*, 2021.

[153] M. Ren *et al.*, "Learning to reweight examples for robust deep learning," in *ICML*, 2018, pp. 4334–4343.

[154] J. Shu *et al.*, "Meta-Weight-Net: Learning an explicit mapping for sample weighting," in *NeurIPS*, 2019, pp. 1917–1928.

[155] S. Li *et al.*, "Meta-reweighted regularization for unsupervised domain adaptation," *IEEE TKDE*, vol. 35, no. 3, pp. 2781–2795, 2023.

[156] Y. Ge *et al.*, "Automated data denoising for recommendation," *arXiv:2305.07070*, 2023.

[157] T. Wang *et al.*, "Dataset distillation," *arXiv:1811.10959*, 2018.

[158] S. Lei *et al.*, "A comprehensive survey of dataset distillation," *arXiv:2301.05603*, 2023.

[159] N. Sachdeva *et al.*, "Data distillation: A survey," *arXiv:2301.04272v1*, 2023.

[160] N. Loo *et al.*, "Efficient dataset distillation using random feature approximation," in *NeurIPS*, 2022.

[161] B. Zhao *et al.*, "Dataset condensation with gradient matching," in *ICLR*, 2021.

[162] J.-H. Kim *et al.*, "Dataset condensation via efficient synthetic-data parameterization," in *ICML*, 2022.

[163] G. Cazenavette *et al.*, "Dataset distillation by matching training trajectories," in *CVPR*, 2022.

[164] J. Cui *et al.*, "Scaling up dataset distillation to imagenet-1k with constant memory," *arXiv:2211.10586*, 2022.

[165] K. Wang *et al.*, "Cafe: Learning to condense dataset by aligning features," in *CVPR*, 2022, pp. 12 196–12 205.

[166] X. Zhou *et al.*, "Probabilistic bilevel coreset selection," in *ICML*, 2022, pp. 27 287–27 302.

[167] K. Meding *et al.*, "Trivial or impossible – dichotomous data difficulty masks model differences (on imagenet and beyond)," in *ICLR*, 2022.

[168] V. Feldman *et al.*, "What neural networks memorize and why: Discovering the long tail via influence estimation," in *NeurIPS*, 2020, pp. 2881–2891.

[169] C. G. Northcutt *et al.*, "Learning with confident examples: Rank pruning for robust classification with noisy labels," *arXiv:1705.01936*, 2017.

[170] Y. Yang *et al.*, "Towards sustainable learning: Coresets for data-efficient deep learning," in *ICML*, 2023.

[171] B. Mirzasoleiman *et al.*, "Coresets for data-efficient training of machine learning models," in *ICML*, 2020, pp. 6950–6960.

[172] Z. Liu *et al.*, "Divaug: Plug-in automated data augmentation with explicit diversity maximization," in *ICCV*, 2021, pp. 4762–4770.

[173] W. Li *et al.*, "Regularization via structural label smoothing," in *AISTATS*, 2020, pp. 1453–1463.

[174] C. Meister *et al.*, "Generalized entropy regularization or: There's nothing special about label smoothing," in *ACL*, 2020.

[175] J. Chai *et al.*, "Fairness with adaptive weights," in *ICML*, 2022, pp. 2853–2866.

[176] S. Hu *et al.*, "Tkml-ap: Adversarial attacks to top-k multi-label learning," in *ICCV*, 2021, pp. 7649–7657.

[177] H. Cao *et al.*, "Mitigating exposure bias in grammatical error correction with data augmentation and reweighting," in *EACL*, 2023, pp. 2123–2135.

[178] Y. Zhao *et al.*, "Adaptive logit adjustment loss for long-tailed visual recognition," in *AAAI*, 2022, pp. 3472–3480.

[179] Y. Wang *et al.*, "Do generated data always help contrastive learning?" in *ICLR*, 2024.

[180] I. Shumailov *et al.*, "Ai models collapse when trained on recursively generated data," *Nature*, vol. 631, p. 755–759, 2024.

[181] S. Hu *et al.*, "A survey on information bottleneck," *IEEE TPAMI*, vol. 46, no. 8, pp. 5325–5344, 2024.

[182] J.-H. Xue *et al.*, "Why does rebalancing class-unbalanced data improve auc for linear discriminant analysis?" *IEEE TPAMI*, vol. 37, no. 5, pp. 1109–1112, 2015.

[183] C. Zheng *et al.*, "Toward understanding generative data augmentation," in *NeurIPS*, 2023, pp. 54 046–54 060.

[184] H. Liu *et al.*, "Self-supervised learning is more robust to dataset imbalance," *arXiv:2110.05025*, 2022.

[185] Q. Dong *et al.*, "Imbalanced deep learning by minority class incremental rectification," *IEEE TPAMI*, vol. 41, no. 6, pp. 1367–1381, 2019.

[186] K. A. Wang *et al.*, "Is importance weighting incompatible with interpolating classifiers?" in *ICLR*, 2022.

[187] D. Xu *et al.*, "Understanding the role of importance weighting for deep learning," *arXiv:2103.15209*, 2021.

[188] R. Xu *et al.*, "A theoretical analysis on independence-driven importance weighting for covariate-shift generalization," in *ICML*, 2022, pp. 24 803–24 829.

[189] D. Chen *et al.*, "Zero-shot logit adjustment," in *IJCAI*, 2022.

[190] M. Qraitem *et al.*, "Bias mimicking: A simple sampling approach for bias mitigation," in *CVPR*, 2023, pp. 20 311–20 320.

[191] Y. Roh *et al.*, "Sample selection for fair and robust training," in *NeurIPS*, 2021, pp. 815–827.

[192] A. Zhang *et al.*, "Boosting causal discovery via adaptive sample reweighting," in *ICLR*, 2023.

[193] Y. Jang *et al.*, "Adversarial defense via learning to generate diverse attacks," in *ICCV*, 2019.

[194] I. Hounie *et al.*, "Automatic data augmentation via invariance-constrained learning," in *ICML*, 2023, pp. 13 410–13 433.

[195] T. Doan *et al.*, "A theoretical analysis of catastrophic forgetting through the ntk overlap matrix," in *AISTATS*, 2021, pp. 1072–1080.

[196] S. Chatterjee *et al.*, "On the generalization mystery in deep learning," *arXiv:2203.10036*, 2022.

[197] Z. Wang *et al.*, "Less is better: Unweighted data subsampling via influence function," in *AAAI*, 2020, pp. 6340–6347.

[198] T. Dao *et al.*, "A kernel theory of modern data augmentation," in *ICML*, 2019, pp. 1528–1537.

[199] J. Wu *et al.*, "A unified framework for adversarial attacks on multi-source domain adaptation," *IEEE TKDE*, pp. 1–12, 2022.

[200] M. Yi *et al.*, "Improved ood generalization via adversarial training and pretraing," in *ICML*, 2021, pp. 11 987–11 997.

[201] T. Fang *et al.*, "Rethinking importance weighting for deep learning under distribution shift," in *NeurIPS*, 2020, pp. 11 996–12 007.

[202] D. Weinshall *et al.*, "Curriculum learning by transfer learning: Theory and experiments with deep networks," in *ICML*, 2018, pp. 5238–5246.

[203] D. Zhu *et al.*, "Rethinking data distillation: Do not overlook calibration," in *ICCV*, 2023.

[204] T. Dong *et al.*, "Privacy for free: How does dataset condensation help privacy?" *arXiv:2206.00240*, 2022.

[205] C. F. G. D. Santos *et al.*, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM CSUR*, vol. 54, no. 10, pp. 1–25, 2022.

[206] L. Yuan *et al.*, "Revisiting knowledge distillation via label smoothing regularization," in *CVPR*, 2020.

[207] A. D. Assis *et al.*, "Neural networks regularization with graph-based local resampling," *IEEE Access*, vol. 9, pp. 50 727–50 737, 2021.

[208] J. Yoon *et al.*, "Data valuation using reinforcement learning," in *ICML*, 2020, pp. 10 842–10 851.

[209] K. Nishi *et al.*, "Augmentation strategies for learning with noisy labels," in *CVPR*, 2021, pp. 8022–8031.

[210] F. R. Cordeiro *et al.*, "Propmix: Hard sample filtering and proportional mixup for learning with noisy labels," in *BMVC*, 2021.

[211] K. Yang *et al.*, "Adversarial auto-augment with label preservation: A representation learning principle guided approach," in *NeurIPS*, 2022, pp. 22 035–22 048.

[212] J. Li *et al.*, "Dividemix: Learning with noisy labels as semi-supervised learning," in *ICLR*, 2020.

[213] J. Shu *et al.*, "Cmw-net: Learning a class-aware sample weighting mapping for robust deep learning," *IEEE TPAMI*, vol. 45, no. 10, 2023.

[214] Z. Zhang *et al.*, "Learning fast sample re-weighting without reward data," in *ICCV*, 2021, pp. 725–734.

[215] X. Wang *et al.*, "Derivative manipulation for general example weighting," *arXiv:1905.11233*, 2020.

[216] E. Yang *et al.*, "Distillhash: Unsupervised deep hashing by distilling data pairs," in *CVPR*, 2019, pp. 2946–2955.

[217] B. Mirzasoleiman *et al.*, "Coresets for robust training of deep neural networks against noisy labels," in *NeurIPS*, 2020, pp. 11 465–11 477.

[218] S. Mindermann *et al.*, "Prioritized training on points that are learnable, worth learning, and not yet learnt," in *ICML*, 2022, pp. 15 630–15 649.

[219] C. Bellinger *et al.*, "Remix: Calibrated resampling for class imbalance in deep learning," *arXiv:2012.02312*, 2020.

[220] F. Du *et al.*, "Global and local mixture consistency cumulative learning for long-tailed visual recognitions," in *CVPR*, June 2023, pp. 15 814–15 823.

[221] O. Pooladzandi *et al.*, "Adaptive second order coresets for data-efficient machine learning," in *ICML*, 2022, pp. 17 848–17 869.

[222] K. Cao *et al.*, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019, pp. 1567–1578.

[223] G. Zhao *et al.*, "Improved distribution matching for dataset condensation," in *CVPR*, 2023, pp. 7856–7865.

[224] X. Peng *et al.*, "Mixgradient: A gradient-based re-weighting scheme with mixup for imbalanced data streams," *Neural Networks*, vol. 161, pp. 525–534, 2023.

[225] C. Huang *et al.*, "Generative dataset distillation," in *BigCom*, 2021, pp. 212–218.

[226] A. Krizhevsky, "Learning multiple layers of features from tiny images." MIT, 2009.

[227] T. Xiao *et al.*, "Learning from massive noisy labeled data for image classification," in *CVPR*, 2015, p. 2691–2699.

[228] W. Li *et al.*, "Webvision database: Visual learning and understanding from web data," *arXiv:1708.02862*, 2017.

[229] G. Van Horn *et al.*, "The inaturalist species classification and detection dataset," in *CVPR*, 2018, pp. 8769–8778.

[230] Z. Liu *et al.*, "Largescale long-tailed recognition in an open world," in *CVPR*, 2019, p. 2537–2546.

[231] S. P. Karimireddy *et al.*, "Byzantine-robust learning on heterogeneous datasets via bucketing," *arXiv:2006.09365*, 2022.

[232] N. Tsilivis *et al.*, "Can we achieve robustness from data alone?" *arXiv:2006.09365*, 2022.

[233] Z. Liu *et al.*, "Mesa: Boost ensemble imbalanced learning with meta-sampler," in *NeurIPS*, 2020, pp. 14 463–14 474.

[234] Y. Li *et al.*, "Adaptive noisy data augmentation for regularization of undirected graphical models," *arXiv:1810.04851*, 2019.

[235] J. Xiao *et al.*, "Stability analysis and generalization bounds of adversarial training," in *NeurIPS*, 2022, pp. 15 446–15 459.

[236] J. An *et al.*, "Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients," *arXiv:2009.13447*, 2021.

[237] Y. Xu *et al.*, "Towards understanding label smoothing," *arXiv:2006.11653*, 2017.

[238] R. Müller *et al.*, "When does label smoothing help?" in *NeurIPS*, 2019.

[239] B. Chen *et al.*, "An investigation of how label smoothing affects generalization," *arXiv:2010.12648*, 2020.

[240] A. Ilyas *et al.*, "Adversarial examples are not bugs, they are features," in *NeurIPS*, 2019.

[241] Z. Qian *et al.*, "A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies," *Pattern Recognition*, vol. 131, p. 108889, 2022.

[242] K. H. R. Chan *et al.*, "Redunet: a white-box deep network from the principle of maximizing rate reduction," *Journal of Machine Learning Research*, vol. 23, no. 1, pp. 4907–5009, 2022.

[243] Y. Wen *et al.*, "Combining ensembles and data augmentation can harm your calibration," in *ICLR*, 2021.

[244] M. Du *et al.*, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, pp. 68–77, 2019.

[245] C. R. Pochimireddy *et al.*, "Can perceptual guidance lead to semantically explainable adversarial perturbations?" *IEEE J-STSP*, pp. 1–11, 2023.

[246] G. Zelaya *et al.*, "Towards explaining the effects of data preprocessing on machine learning," in *ICDE*, 2019, pp. 2086–2090.

[247] E. Mosqueira-Rey *et al.*, "Human-in-the-loop machine learning: a state of the art," *Journal of Machine Learning Research*, vol. 56, pp. 3005–3054, 2023.

[248] K. M. Collins *et al.*, "Human-in-the-loop mixup," in *UAI*, 2023, pp. 454–464.

[249] E. Wallace *et al.*, "Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering," *TACL*, vol. 7, pp. 387–401, 2019.

[250] Z.-F. Wu *et al.*, "Ngc: a unified framework for learning with open-world noisy data," in *ICCV*, 2021, p. 62–71.

[251] Z. Jiang *et al.*, "Improving contrastive learning on imbalanced data via open-world sampling," in *NeurIPS*, 2021, pp. 5997–6009.

[252] L. Wei *et al.*, "Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4," *arXiv:2308.12067*, 2023.

[253] T. Baltrušaitis *et al.*, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.

[254] Z. Yan *et al.*, "Dehib: Deep hidden backdoor attack on semi-supervised learning via adversarial perturbation," in *AAAI*, 2021, pp. 10 585–10 593.

[255] W. Liang *et al.*, "Advances, challenges and opportunities in creating data for trustworthy ai," *Nature Machine Intelligence*, vol. 4, p. 669–677, 2022.

[256] R. Yu *et al.*, "Dataset distillation: A comprehensive review," *IEEE TPAMI*, vol. 46, no. 1, pp. 150–170, 2024.

[257] D. Yu *et al.*, "How does data augmentation affect privacy in machine learning?" in *AAAI*, 2021, pp. 10 746–10 753.

[258] X. Li *et al.*, "On the privacy effect of data enhancement via the lens of memorization," *IEEE TIFS*, vol. 19, pp. 4686–4699, 2024.

Ou Wu received the B.Sc. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2003, and the M.Sc. and Ph.D. degrees in computer science from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2006 and 2012, respectively. In 2007, he joined NLPR as an Assistant Professor. In 2017, he joined Center for Applied Mathematics, Tianjin University, China, as a full professor. In 2024, he became a full professor in Hangzhou Insititute of Advanced Study (HIAS), University of Chinese Acadamy of Sciences, China. His research interests include AI data optimization, synthesis, and security.

Rujing Yao received the B.Sc. degree in information and computing science from Jilin Agricultural University, Jilin, China, in 2018, and the M.Sc. degree in mathematics from the Center for Applied Mathematics, Tianjin University, Tianjin, China, in 2021. She is currently working toward the Ph.D. degree with the Department of Information Resources Management, Business School, Nankai University, Tianjin, China. Her research interests include AI data optimization and data mining.

Supplementary Materials to "Data Optimization for Deep Learning: A Survey"

I. The sub-taxonomy for data optimization techniques with methods in each category

To facilitate interesting readers to better understand our proposed sub-taxonomy for data optimization techniques, the structure of our proposed sub-taxonomy as well as several methods in each category is shown in Fig. 1. The methods (i.e., their corresponding references) with the red color are directly from the paper and the ones (i.e., their corresponding references) [1]–[43] with the green color only appear in this file. There methods were published in the past two years.

II. More details for some methods in Section VI

Section VI in the paper introduces a number of typical (not exhaustive) methods in each technical path for data optimization. Details for some methods are presented in this section due to lack of space for the paper.

RISDA [44]: This method is an improved version of ISDA [45]. When application ISDA, the calculation of mean vectors covariance matrix of some categories with small numbers of samples may be unreliable. Therefore, RISDA enriches the mean vectors and covariance matrices of these categories by introducing the covariance matrices of their similar categories.

Label smoothing [46]. It is a kind of label perturbation method. Let C be the number of categories and λ be a hyper-parameter. Label smoothing perturbs the label y (one-hot type) with the following perturbation $\triangle y = \lambda(\frac{I}{C} - y)$, where I is a C-dimensional vector and its each element is 1.

Universal adversarial perturbation [47]: Most adversarial perturbation methods yield different perturbations for different samples. Nevertheless, Shafahi et al. [47] simplified the optimization approach by restricting that all the involved samples share the same perturbation, which is called universal adversarial perturbation (UAP). Extensive experiments indicate that such a simple strategy is useful as the whole time consumption is significantly reduced while UAP is also beneficial for robust training.

Class-wise adversarial perturbation [48]: Different from UAP requiring that all training samples share the identical perturbation, class-wise adversarial perturbation requires that training samples in a category share the same perturbation. This relaxation is also beneficial for the entire training.

Class-wise logit perturbation [49]: This method is designed to balance feature distribution. It defines category-wise logit perturbation for samples in each category with the manner that head categories are assigned with smaller perturbation while tail ones are assigned with larger perturbation.

Sparse-regularized perturbation [50]: Wei et al. [50] investigated the adversarial attach for videos. The found that not all frames in a video requires extra perturbations as there are temporal interactions among frames. Therefore, they designed an $l2$,1-norm based optimization algorithm to compute the sparse adversarial perturbations for videos.

Self-supervised perturbation [51]: Convectional adversarial perturbation relies on the supervised labels. Naseer et al. [51] investigated the pursue of the adversarial perturbation without requiring labels. Their method is based on maximization of feature distortion for each training sample.

Curriculum learning [52]: It is an independent learning paradigm proposed by Bengio et al. [52]. It mimics the human learning procedure, advocating for models to start learning from easy samples and gradually progress to hard samples. This paradigm can improve the generalization ability and convergence rate of various DNN models in various learning tasks in computer vision and natural language processing.

Self-paced learning [53]: This strategy can be seen as in concrete implementation of curriculum learning. It takes the training loss as the indicator of the learning difficulty of samples. Samples with losses smaller than a threshold are considered easy. In an epoch, only easy samples are allowed to take participant in training and the threshold grows gradually at each epoch.

Importance weighting [54]: This procedure assign large weights to the training samples that are more likely to appear in the test data and small weights to those that are less likely. The weighted training loss can thus be minimized by conventional learning algorithms, resulting in a simple and general scheme to deal with learning tasks with distribution shift.

Meta learning-based weighting: Ren et al. [55] initially introduced meta learning into the pursuing of sample weights in deep learning. This line of technique takes the sample weights as parameters to learn. A meta dataset (or validation dataset) is utilize to learn the weights in each training epoch. Therefore, there are two optimization procedures in each training epoch and thus the entire time consumption is higher than many other weighting methods.

III. Another Future Direction: Data optimization agent

Given a concrete learning task, a selection dilemma occurs for the tremendous data optimization techniques. There have been studies on the automatic data optimization such as automatic data augmentation [56]. Nevertheless, existing automatic data optimization methods still focus on a particular type of technical path rather than the types across different technical paths [57], [58]. A more general data optimization agent can be trained by iteratively training on a large number of deep learning tasks via reinforcement learning.

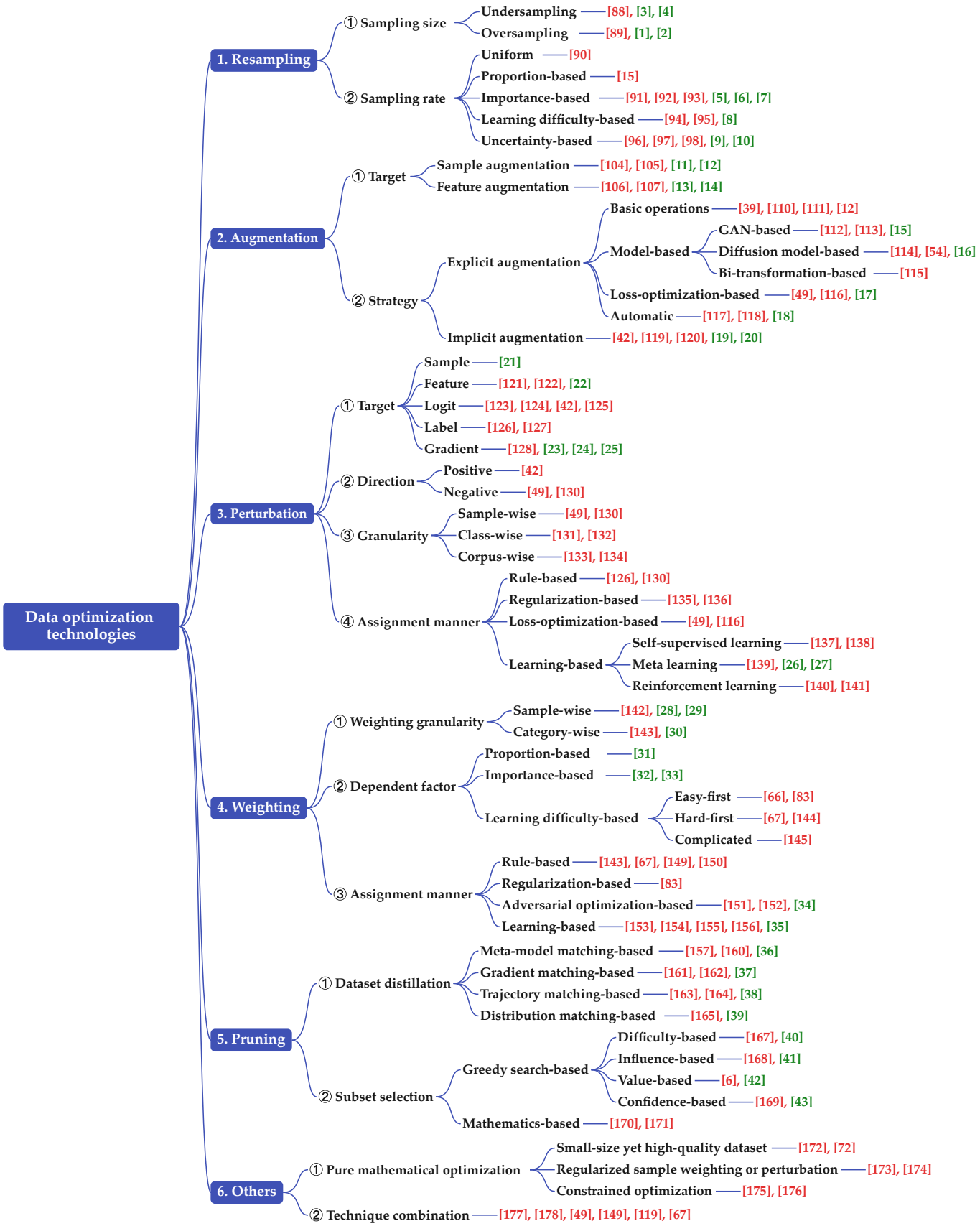

Fig. 1. The sub-taxonomy of data optimization techniques with categorized methods.

REFERENCES

[1] T. Li, H. Xu, W. Tan, K. Murray, and D. Khashabi, "Upsample or upweight? balanced training on heavily imbalanced datasets," 2024.

[2] K. Yang, Z. Yu, W. Chen, Z. Liang, and C. L. P. Chen, "Solving the imbalanced problem by metric learning and oversampling," *IEEE TKDE*, pp. 1–14, 2024.

[3] Z. Sun, W. Ying, W. Zhang, and S. Gong, "Undersampling method based on minority class density for imbalanced data," *Expert Systems with Applications*, vol. 249, no. 1, 2024.

[4] H. Yu, Y. Du, and J. Wu, "Reviving undersampling for long-tailed learning," 2024.

[5] H. Hajimolahoseini, O. M. Awad, W. Ahmed, A. Wen, S. Asani, M. Hassanpour, F. Javadi, M. Ahmadi, F. Ataiefard, K. Liu, and Y. Liu, "Swiftlearn: A data-efficient training method of deep learning models using importance sampling," 2023.

[6] A. Li, Y. Chen, C. Ren, W. Wang, M. Hu, T. Li, H. Yu, and Q. Chen, "Federated graph learning with adaptive importance-based sampling," 2024.

[7] S. Lu, Y. Hu, L. Yang, Z. Sun, J. Mei, J. Tan, and C. Song, "Pa&da: Jointly sampling path and data for consistent nas," in *CVPR*, June 2023, pp. 11 940–11 949.

[8] T. Jang and X. Wang, "Difficulty-based sampling for debiased contrastive representation learning," in *CVPR*, June 2023, pp. 24 039–24 048.

[9] A. Hoarau, V. Lemaire, Y. L. Gall, J.-C. Dubois, and A. Martin, "Evidential uncertainty sampling strategies for active learning," *Machine Learning*, vol. 113, p. 6453–6474, 2024.

[10] S. Ma, H. Wu, A. Lawlor, and R. Dong, "Breaking the barrier: Selective uncertainty-based active learning for medical image segmentation," in *ICASSP*, 2024, pp. 1531–1535.

[11] F. Iqbal, A. Abbasi, A. R. Javed, A. Almadhor, Z. Jalil, S. Anwar, and I. Rida, "Data augmentation-based novel deep learning method for deepfaked images detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 11, 2024.

[12] "A novel data augmentation approach to fault diagnosis with class-imbalance problem," *Reliability Engineering & System Safety*, vol. 243, p. 109832, 2024.

[13] Q. Wang, J. Jia, Y. Deng, J. Chen, X. Wang, M. Huang, and A. H. Aghvami, "Darloc: Deep learning and data-feature augmentation based robust magnetic indoor localization," *Expert Systems with Applications*, vol. 244, p. 122921, 2024.

[14] T. Zhou, Y. Yuan, B. Wang, and E. Konukoglu, "Federated feature augmentation and alignment," *IEEE TPAMI*, pp. 1–17, 2024.

[15] A. Kumar and D. Singh, "Generative adversarial network-based augmentation with noval 2-step authentication for anti-coronavirus peptide prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–13, 2024.

[16] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, and Z. Liu, "Remodiffuse: Retrieval-augmented motion diffusion model," in *ICCV*, October 2023, pp. 364–373.

[17] W. Sun, H. Wang, and R. Qu, "A novel data generation and quantitative characterization method of motor static eccentricity with adversarial network," *IEEE Transactions on Power Electronics*, vol. 38, no. 7, pp. 8027–8032, 2023.

[18] D. Tomar, G. Vray, B. Bozorgtabar, and J.-P. Thiran, "Tesla: Test-time self-learning with automatic adversarial augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 20 341–20 350.

[19] W. Zhu, O. Wu, and N. Yang, "Irda: Implicit data augmentation for deep imbalanced regression," *Information Sciences*, vol. 677, p. 120873, 2024.

[20] H. Feng, J. Yan, J. Liu, J. Zheng, and Q. Ma, "Well begun is half done: An implicitly augmented generative framework with distribution modification for hierarchical text classification," in *LREC-COLING 2024*, 2024, pp. 17 433–17 443.

[21] S. Zhang, F. Liu, J. Yang, Y. Yang, C. Li, B. Han, and M. Tan, "Detecting adversarial data by probing multiple perturbations using expected perturbation score," in *ICML*, 2023, pp. 41 429–41 451.

[22] S. Udupa, P. Gurunath, A. Sikdar, and S. Sundaram, "Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation," in *CVPR*, 2024, pp. 5904–5914.

[23] C. Feng, N. Xu, W. Wen, P. Venkitasubramaniam, and C. Ding, "Spectral-dp: Differentially private deep learning through spectral perturbation and filtering," in *2023 IEEE Symposium on Security and Privacy (SP)*, 2023, pp. 1944–1960.

[24] L. Chen, D. Yue, X. Ding, Z. Wang, K.-K. R. Choo, and H. Jin, "Differentially private deep learning with dynamic privacy budget allocation and adaptive optimization," *IEEE TIFS*, vol. 18, pp. 4422–4435, 2023.

[25] M. Gogoi, S. Tiwari, and S. Verma, "Perturbing the gradient for alleviating meta overfitting," 2024.

[26] F. Yin, Y. Zhang, B. Wu, Y. Feng, J. Zhang, Y. Fan, and Y. Yang, "Generalizable black-box adversarial attack with meta learning," *IEEE TPAMI*, vol. 46, no. 3, pp. 1804–1818, 2024.

[27] C. Yu, Z. Zhang, H. Li, J. Sun, and Z. Xu, "Meta-learning-based adversarial training for deep 3d face recognition on point clouds," *Pattern Recognition*, vol. 134, p. 109065, 2023.

[28] W. Jiang, T. Chen, G. Ye, W. Zhang, L. Cui, Z. Huang, and H. Yin, "Physics-guided active sample reweighting for urban flow prediction," 2024.

[29] H. Yang, M. Wang, Z. Yu, H. Zhang, J. Jiang, and Y. Zhou, "Confidence-based and sample-reweighted test-time adaptation," *Knowledge-Based Systems*, vol. 283, p. 111164, 2024.

[30] X. Chen, Y. Zhou, D. Wu, C. Yang, B. Li, Q. Hu, and W. Wang, "Area: Adaptive reweighting via effective area for long-tailed classification," in *ICCV*, 2023, pp. 19 277–19 287.

[31] S. Guan, X. Zhao, Y. Xue, and H. Pan, "Awgan: An adaptive weighting gan approach for oversampling imbalanced datasets," *Information Sciences*, vol. 663, p. 120311, 2024.

[32] B. He and C. Ma, "Interpretable triplet importance for personalized ranking," 2024.

[33] M. Drnevich, S. Jiggins, J. Katzy, and K. Cranmer, "Neural quasiprobabilistic likelihood ratio estimation with negatively weighted data," 2024.

[34] X. Gu, X. Yu, Y. Yang, J. Sun, and Z. Xu, "Adversarial reweighting with α -power maximization for domain adaptation," *International Journal of Computer Vision*, vol. 132, p. 4768–4791, 2024.

[35] Z. Chen, T. Xiao, K. Kuang, Z. Lv, M. Zhang, J. Yang, C. Lu, H. Yang, and F. Wu, "Learning to reweight for generalizable graph neural network," in *AAAI*, 2024, pp. 8320–8328.

[36] S. Liu and X. Wang, "Mgdd: A meta generator for fast dataset distillation," in *NeurIPS*, 2023, pp. 56 437–56 455.

[37] C. Wang, J. Sun, Z. Dong, R. Li, and R. Zhang, "Gradient matching for categorical data distillation in ctr prediction," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023, p. 161–170.

[38] Y. Lee and H. W. Chung, "Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching," in *ICML*, 2024.

[39] G. Zhao, G. Li, Y. Qin, and Y. Yu, "Improved distribution matching for dataset condensation," in *CVPR*, 2023, pp. 7856–7865.

[40] A. Acharya, D. Yu, Q. Yu, and X. Liu, "BOSS: Diversity-difficulty balanced one-shot subset selection for data-efficient deep learning," 2024. [Online]. Available: https://openreview.net/forum?id=QcgvtqxRhI

[41] T. Wan, K. Xu, L. Lan, Z. Gao, F. Dawei, B. Ding, and H. Wang, "Tracing training progress: Dynamic influence based selection for active learning," in *ACM Multimedia*, 2024.

[42] Y. He, Z. Wang, Z. Shen, G. Sun, Y. Dai, Y. Wu, H. Wang, and A. Li, "Shed: Shapley-based automated dataset refinement for instruction fine-tuning," 2024.

[43] Y. Kong, L. Liu, M. Qiao, Z. Wang, and D. Tao, "Trust-region adaptive frequency for online continual learning," *International Journal of Computer Vision*, vol. 131, p. 1825–1839, 2023.

[44] X. Chen, Y. Zhou, D. Wu, W. Zhang, Y. Zhou, B. Li, and W. Wang, "Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification," in *AAAI*, Online, February 2022, pp. 356–364.

[45] Y. Wang, X. Pan, S. Song, H. Zhang, C. Wu, and G. Huang, "Implicit semantic data augmentation for deep networks," in *NeurIPS*, 2019, pp. 12 635–12 644.

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[47] A. Shafahi, M. Najibi, Z. Xu, J. Dickerson, L. S. Davis, and T. Goldstein, "Universal adversarial training," in *CVPR*, 2017, pp. 5636–5643.

[48] P. Benz, C. Zhang, A. Karjauv, and I. S. Kweon, "Universal adversarial training with class-wise perturbations," in *ICME*, 2021, pp. 1–6.

[49] Y. Wang, J. Fei, H. Wang, W. Li, T. Bao, L. Wu, R. Zhao, and Y. Shen, "Balancing logit variation for long-tailed semantic segmentation," in *CVPR*, 2023, pp. 19 561–19 573.

[50] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse adversarial perturbations for videos," in *AAAI*, 2019, pp. 8973–8980.

[51] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *CVPR*, 2020, pp. 262–271.

[52] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *ICML*, 2009, pp. 41–48.
[53] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," *NeurIPS*, pp. 1–9, 2010.
[54] T. Liu and D. Tao, "Classification with noisy labels by importance reweighting," *IEEE TPAMI*, vol. 38, no. 3, p. 447–461, 2023.
[55] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *ICML*, 2018, pp. 4334–4343.
[56] I. Hounie, L. F. O. Chamon, and A. Ribeiro, "Automatic data augmentation via invariance-constrained learning," in *ICML*, 2023, pp. 13 410–13 433.
[57] V. A. Trinh, H. Salami Kavaki, and M. I. Mandel, "Importantaug: A data augmentation agent for speech," in *ICASSP*, 2022, pp. 8592–8596.
[58] M. Li, X. Zhang, C. Thrampoulidis, J. Chen, and S. Oymak, "Autobalance: Optimized loss functions for imbalanced data," in *NeurIPS*, 2021, pp. 3163–3177.