Sentinel-Guided Zero-Shot Learning: A Collaborative Paradigm Without Real Data Exposure

Fan Wan, Xingyu Miao, Haoran Duan, *Graduate Student Member, IEEE*, Jingjing Deng, Rui Gao, and Yang Long, *Senior Member, IEEE*

***Abstract*— With increasing concerns over data privacy and model copyrights, especially in the context of collaborations between AI service providers and data owners, an innovative Sentinel-Guided Zero-Shot Learning (SG-ZSL) paradigm is proposed in this work. SG-ZSL is designed to foster efficient collaboration without the need to exchange models or sensitive data. It consists of a teacher model, a student model and a generator that links both model entities. The teacher model serves as a sentinel on behalf of the data owner, replacing real data, to guide the student model at the AI service provider's end during training. Considering the disparity of knowledge space between the teacher and student, we introduce two variants of the teacher model: the omniscient and the quasi-omniscient teachers. Under these teachers' guidance, the student model seeks to match the teacher model's performance and explores domains that the teacher has not covered. To trade-off between privacy and performance, we further introduce two distinct security-level training protocols: white-box and black-box, enhancing the paradigm's adaptability. Despite the inherent challenges of real data absence in the SG-ZSL paradigm, it consistently outperforms in ZSL and GZSL tasks, notably in the white-box protocol. Our comprehensive evaluation further attests to its robustness and efficiency across various setups, including stringent black-box training protocol.**

***Index Terms*— Data-free knowledge transfer, privacy protection, zero-shot learning.**

I. Introduction

THE profound advancements in deep learning can be largely attributed to the evolution of high-performance computing and the proliferation of extensive multimodal datasets. At the heart of deep learning lies the ability of pre-trained models, such as ResNet [1] for visual features and BERT [2] for semantic nuances, to distill empirical knowledge from vast datasets. Through these models, intricate patterns and relationships within expansive data terrains are effectively discerned, proving pivotal for addressing intricate real-world challenges. Nonetheless, the sharing of data among institutions has increasingly been met with complexities and apprehensions. The broader public's concerns regarding data ownership, copyright implications, the financial burden of large-scale annotations, and restricted access to domain-specific data have hindered the progression of interdisciplinary and intercultural deep learning models. Furthermore, the hesitation to share models publicly often comes from concerns about intellectual property, misuse, and protecting private and sensitive information. Balancing the desire for openness with these issues is an ongoing struggle for researchers.

As depicted in Fig.1, datasets, which may encompass sensitive entities such as personal healthcare records and facial images, have necessitated substantial financial and temporal investments from data proprietors. Stringent regulations, epitomized by Europe's GDPR and the California Consumer Privacy Act (CCPA), have been instituted to safeguard personal data and uphold user privacy. As a result, the acquisition, transmission, and dissemination of such data have become increasingly intricate and laden with challenges. Federated Learning (FL) [3], as a pioneering privacy-preserving paradigm, offers a decentralized approach to model training. This method facilitates the training of a global shared model across multiple parties through the mere exchange of model updates without centralizing data, which largely preserves data privacy. Nonetheless, issues coexist, including ambiguities in model ownership, potential disclosure of proprietary information, uncertainties in intellectual property rights, and threats stemming from unauthorized model deployment. The fact that many AI service providers are unwilling to disclose the training datasets and internal parameters of Large Language Models is an illustration of these dilemmas. Given the mutual dependencies between AI service providers (i.e., AI companies) and data owners (i.e., research institutions and hospitals), direct data or model sharing poses potential infringements on copyright and privacy mandates. This situation underscores the critical demand for inventive approaches that enable collaboration while safeguarding sensitive information and adhering to copyright laws. Against this backdrop of challenges, our study embarks

Manuscript received 20 November 2023; revised 16 February 2024; accepted 26 March 2024. Date of publication 4 April 2024; date of current version 30 September 2024. This work was supported in part by the U.K. Medical Research Council (MRC) Innovation Fellowship under Grant MR/S003916/2 and in part by the International Exchanges under Grant 2022 IEC\NSFC\223523. This article was recommended by Associate Editor Y. S. Rawat. *(Fan Wan, Rui Gao, and Haoran Duan contributed equally to this work.) (Corresponding author: Yang Long.)*

Fan Wan, Xingyu Miao, Haoran Duan, Jingjing Deng, and Yang Long are with the Department of Computer Science, Durham University, DH1 3LE Durham, U.K. (e-mail: fan.wan@durham.ac.uk; xingyu.miao@durham.ac.uk; haoran.duan@durham.ac.uk; jingjing.deng@durham.ac.uk; yang.long@durham.ac.uk).

Rui Gao is with the School of Computing and Mathematic Sciences, University of Leicester, LE1 7RH Leicester, U.K. (e-mail: rg344@leicester.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3384756.

Digital Object Identifier 10.1109/TCSVT.2024.3384756

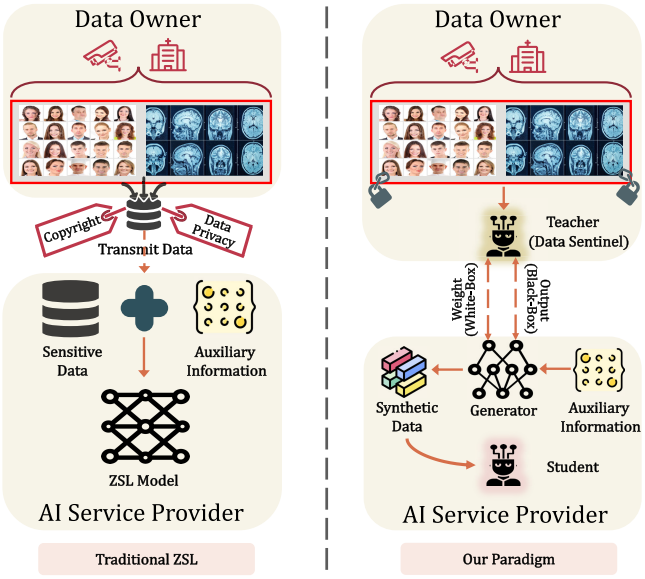


Fig. 1. In traditional ZSL approaches, real data is necessitated to establish the visual-semantic association. Conversely, SG-ZSL introduces a teacher model, which acts as a data sentinel, enabling the execution of ZSL tasks without the need for direct access to real data.

on an exploration within a strict framework where neither real data nor models are disclosed to the public. We endeavor to elucidate a collaborative learning paradigm necessitating no real data exchange, wherein knowledge transfer transpires through a teacher-student distillation facilitated by a data-generative mechanism.

Zero-shot learning (ZSL), an emerging machine learning paradigm, demonstrates considerable potential in addressing data-free challenges, particularly in scenarios where deep transfer extends beyond the seen classes present in the training dataset. By establishing a robust visual-semantic linkage through auxiliary information, such as attributes or word embeddings, ZSL empowers models to recognize unseen classes without prior exposure to relevant data. Nevertheless, conventional ZSL models predominantly depend on actual data from either seen or unseen categories. In adapting pre-trained models to novel task domains, an implicit assumption is made regarding the presence of a significant volume of labeled seen class or unlabeled unseen class data to build the visual-semantic connection. This assumption, however, is often contradicted to the restrictive nature of data sharing across varied institutions and countries.

In this paper, we introduce SG-ZSL, as shown in Fig. 1. The proposed approach markedly diverges from traditional ZSL, aiming to prevent the leakage of sensitive data while still enabling effective training of AI models. Within this framework, the teacher model, pre-trained on real data, assumes the dual responsibility of data protection and guidance for the ZSL model located at the AI Service Provider's end, thereby ensuring model training without direct exposure to the original data from the Data Owner. To reconcile privacy concerns with performance objectives, two distinct SG-ZSL protocols are presented: a 'black-box' version, which solely conveys the output classification scores from the teacher model and a 'white-box' variant that shares both model weights and classification scores. These protocols are tailored to meet the diverse needs of data owners and AI service providers concerning model efficacy and data security. To bolster defenses against potential threats, Differential Privacy (DP) [4] is integrated into our teacher model's training process. The direct transfer of the teacher model to the AI Service Provider is prohibited; instead, the latter must submit specific requests to receive training guidance from the teacher model. This mechanism allows the Data Owner to exert control over data security by regulating request frequency. Recognizing potential disparities in the knowledge domains of the teacher and student models, our teacher models are further categorized into omniscient and quasi-omniscient types, distinguished primarily based on their exposure to unseen categories during the pre-training phase.

In summary, our contributions encompass:

- We introduce Sentinel-Guided Zero-Shot Learning, a novel paradigm for Zero-Shot classification without real data access, addressing crucial data privacy and model copyright issues.
- To meet the multifaceted needs in terms of privacy preservation and performance optimization, we formulate two distinct training protocols: white-box and black-box. Additionally, we analyze teacher models under omniscient and quasi-omniscient scenarios within the knowledge space, enhancing our paradigm's adaptability.
- We showcase experimental results for both conventional and generalized ZSL tasks in two scenarios. Despite the lack of data sharing during training, the SG-ZSL model yields promising performance, highlighting our approach's viability.

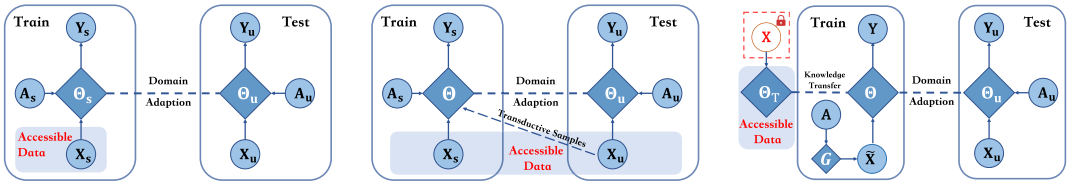
II. Related Work

The realm of machine learning has recently experienced a significant shift towards prioritizing data privacy, particularly when handling sensitive information across diverse domains. Federated Learning [3] has been recognized as a formidable framework, designed to mitigate potential data leakage by decentralizing the training process. Recent advancements in this domain have been characterized by the exploration of various architectures and optimization strategies, all aimed at enhancing model performance without sacrificing data privacy. For example, studies [5], [6], [7], [8] have been dedicated to optimizing communication efficiency in federated learning setups, while research such as [9] and [10] has delved into the application of federated learning in edge computing, ensuring data privacy at its source.

Differential Privacy [4] has been seamlessly integrated into numerous machine learning paradigms to bolster data privacy. Recent contributions, including [11], [12], [13], have investigated the fusion of DP with deep learning, ensuring that while models remain proficient, the privacy of their training data remains uncompromised. For example, Guo et al. [11] developed 'TOP-DP', a topology-aware differential privacy approach for decentralized image classification systems, which innovatively utilizes decentralized communication topologies to enhance privacy protection while achieving an improved balance between model usability and data privacy.

TABLE I

THE DISTINCTIONS BETWEEN SG-ZSL AND TRADITIONAL ZSL SETTINGS ARE DELINEATED IN THE TABLE. HEREIN, 'S' AND 'U' DENOTE THE SEEN AND UNSEEN CLASSES, RESPECTIVELY. '$\mathcal{X}$' SIGNIFIES VISUAL FEATURES, WHILE '$\tilde{\mathcal{X}}$' PERTAINS TO GENERATED FEATURES. THE SEMANTICS OF THE SEEN AND UNSEEN CLASSES ARE REPRESENTED BY 'A_s' AND 'A_u', RESPECTIVELY. THE RED 'X' SYMBOLIZES SENSITIVE REAL DATA. THE ZSL MODEL IS DENOTED BY 'θ', WHEREAS 'θ_T' CORRESPONDS TO THE PRE-TRAINED TEACHER MODEL SPECIFIC TO THE SG-ZSL TASK. 'θ_U' CAN BE ASSOCIATED WITH EITHER THE CONVENTIONAL ZSL MODEL OR THE SG-ZSL MODEL. IT SHOULD BE NOTED THAT THE SG-ZSL MODEL IS CONSTRUCTED UNDER THE GUIDANCE OF THE TEACHER MODEL, EFFECTIVELY ELIMINATING THE NEED FOR SHARING ACTUAL DATA

	IZSL	TZSL	SG-ZSL
Accessible Data	$\mathcal{X}_s$	$\mathcal{X}_s \cup \mathcal{X}_u$	**0**
Accessible Weights	θ_S	θ_{S+U}	θ_T / **0**
Paradigm			

Knowledge Distillation [14], on the other hand, has emerged as a pivotal strategy for protecting intricate teacher models by training a streamlined student model, thereby thwarting potential adversarial attacks. Recent endeavors, such as [15], [16], [17], and [18], have showcased the versatility of knowledge distillation across domains of computer vision. For example, Zhang et al. [18] introduced an evolutionary knowledge distillation approach, where an adaptive, online-evolving teacher model continuously transfers intermediate knowledge to a student network, significantly enhancing learning effectiveness, especially in low-resolution and few-sample scenarios.

It is imperative to note, however, that both Federated Learning and Knowledge Distillation are predominantly confined to supervised learning. This confines their utility in scenarios necessitating the recognition and categorization of previously unseen data categories, a domain where Zero-Shot Learning protocols excel. ZSL, with its prowess in recognizing unseen classes by establishing semantic relationships, transcends the limitations inherent to the supervised nature of both Federated Learning and Knowledge Distillation.

In this work, an innovative SG-ZSL paradigm is introduced. This paradigm, distinct in its data-free knowledge transfer, is adept at addressing unseen data categories, especially in contexts where data sensitivity and privacy are paramount. The incorporation of DP within the teacher model further enhances data privacy, ensuring that the traditional ZSL generalization properties to unseen classes are preserved without additional training, all while safeguarding data and model privacy.

Zero-Shot Learning [19], [20], [21], [22], [23], [24], [25] is predicated on recognizing unseen classes by establishing connections between seen and unseen classes through semantic information, such as attributes [26], [27], [28], [29], word embeddings [30] and predefined similes [31], [32]. Numerous studies [33], [34], [35] have been dedicated to mapping from visual to semantic space, while others [23], [36], [37], [38] focus on generating unseen class data to mitigate data scarcity issues. Effective spaces for visual and semantic embedding have been investigated in [39], [40], [41], [42], and [43]. Depending on the utilization of unseen data during training, ZSL methods can be categorized into inductive [44], [45] and transductive settings [46], [47]. As for the test phase, conventional ZSL methods [39], [48] operate under the assumption that test data originates exclusively from unseen classes, while Generalized ZSL (GZSL) [49], [50], [51] aims to classify both seen and unseen data into their respective classes.

The distinctions between SG-ZSL and traditional ZSL settings are elucidated in Table I. In terms of data access during training, IZSL and Transductive ZSL (TZSL) access labeled seen data and data from both seen and unseen classes, respectively. In contrast, the SG-ZSL setting operates without direct data access, relying solely on a teacher model, trained on sensitive real data, for guidance (as indicated by the red 'X' in Table I). Concerning model security, weight accessibility refers to the accessibility of weights trained on real data. While ZSL models in both inductive and transductive settings possess accessible weights, the SG-ZSL paradigm introduces a teacher model pre-trained on real data. In assessing teacher weight privacy, we introduce the black-box and white-box protocols. In the white-box protocol, teacher weights are accessible for guidance during SG-ZSL model training, whereas the black-box protocol restricts weight sharing, thereby preserving the privacy of both data and model weights.

III. METHODOLOGY

As depicted in Fig. 1, in scenarios where the Data Owner's sensitive data is inaccessible yet a collaboration with the AI Service Provider is sought to leverage the data's value, the proposed SG-ZSL paradigm emerges as a solution. The Data Owner employs a teacher model, serving as a data sentinel, which guides the AI Service Provider's models in training classifiers without real data access. Recognizing the balance between privacy preservation and performance optimization,

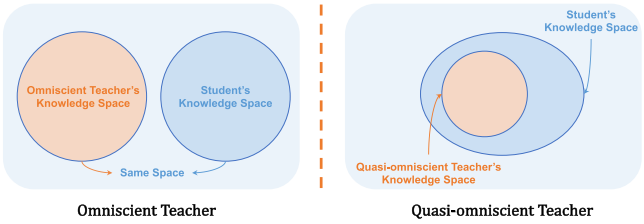


Fig. 2. Differences between the Omniscient and the Quasi-omniscient teacher.

two distinct training protocols with varying security levels, namely the white-box and black-box protocols, are introduced to enhance the paradigm's adaptability.

A. Problem Definition

The SG-ZSL paradigm fosters collaboration between the Data Owner, housing a teacher model, and the AI Service Provider, hosting a student model and a generator. The teacher model, represented as $\mathcal{F}_{\theta_T} : \mathcal{X} \rightarrow \mathcal{Y}$, serves as a data sentinel. Central to the SG-ZSL paradigm is the utilization of the teacher model at the Data Owner's end to direct the training of the student model at the AI Service Provider's end. This objective is achieved through synthetic data generated by the generator $\mathcal{F}_{\theta_G}$, with the aim of enabling the student model to match the teacher's performance or explore domain not covered by the teacher without the transmission of real data. The objective function is given by:

$$\ell\left(\mathcal{F}_{\{\theta_S,\theta_G\}}\left(\tilde{x}\right), \mathcal{F}_{\theta_T}\left(\tilde{x}\right)\right), \tag{1}$$

where ℓ denotes the objective function guided by the teacher, and $\tilde{x} \in \tilde{\mathcal{X}}$ signifies the data generated by the generator, ensuring no real data access.

B. Data Sentinel at the Data Owner's End

1) Omniscient and Quasi-Omniscient Teachers: Given the potential inconsistency between the teacher's data categories and the student model's objective categories, there may be unseen class data absent in the teacher's domain but essential for the student model. Thus, teacher models are further categorized into omniscient and quasi-omniscient types as shown in Fig.2. The omniscient model encompasses all categories, covering both seen and unseen class data, while the quasi-omniscient model is limited to seen class data.

Here, we define the seen class as $\mathcal{S} = \{(x_s, a_s, y_s) \mid x_s \in \mathcal{X}_s, a_s \in \mathcal{A}, y_s \in \mathcal{Y}_s\}$, where $x_s \in \mathbb{R}^{d_x}$ denotes the d_x-dimensional visual feature in the set of seen class features, $a_s \in \mathbb{R}^{d_a}$ denotes the d_a-dimensional auxiliary class-level semantic embedding, and $\mathcal{Y}_s$ stands for the set of labels for seen classes. Unseen classes are defined as $\mathcal{U} = \{(x_u, a_u, y_u) \mid x_u \in \mathcal{X}_u, a_u \in \mathcal{A}, y_u \in \mathcal{Y}_u\}$, where x_u represents the unseen class features, a_u denotes the semantic embedding of unseen classes and y_u denotes the unseen class labels. The seen and unseen classes are disjoint, *i.e.*, $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$.

In the SG-ZSL paradigm, a key constraint is the inaccessibility of both seen and unseen real features at the Data Owner's end during the student model and generator training at the AI Service Provider's end. The available information for the AI service provider is represented as $\mathcal{T}_r = \{(a, y) \mid a \in \mathcal{A}, y \in \mathcal{Y}\}$, indicating only semantic embeddings a and class labels y are available during training. Additionally, a teacher model, pre-trained on real data, is provided to guide the training of the student model and generator. Depending on the teacher model type, different teacher objectives are considered.

2) Teacher Objectives: The teacher models guide the student model in mastering various ZSL tasks. For the CZSL task, the student model's objective is to classify test unseen class images, represented by $f_{ZSL} : \mathcal{X}_u \rightarrow \mathcal{Y}_u$. For the GZSL task, the student model aims to recognize all class images, denoted by $f_{GZSL} : \mathcal{X} \rightarrow \mathcal{Y}$.

3) Incorporating DP in Teacher Model Training: To bolster the protection of sensitive data at the Data Owner's end, differential privacy techniques are seamlessly integrated into the teacher model's training process.

Differential privacy stands as a preeminent mechanism for ensuring data and model security. Denote an algorithm with the differential privacy property by $M(.)$. The algorithm is randomized to make it difficult to have access to the privacy information of the input data. The formal definition of DP is provided below:

Definition 1 [4]: Given a pair of neighboring datasets D and D', for every set of outcomes S, a mechanism M satisfies DP if the following holds:

$$\mathbb{P}(M(D) \in S) \leq e^{\varepsilon} \cdot \mathbb{P}(M(D') \in S) + \delta \tag{2}$$

Here, M(D) and M(D') represent the algorithm's outputs for input datasets D and D', respectively, and $\mathbb{P}$ captures the algorithm's inherent noise randomness. Both ε (privacy budget) and δ (failure probability) influence the privacy strength: smaller values of ε and δ ensure enhanced privacy. In the realm of deep learning, DP is typically realized by introducing the subsampled Gaussian mechanism to safeguard the minibatch gradients during the training process [52], [53], [54]. The distinction between deep learning with DP and conventional deep learning hinges on the private release of the gradient. The Gaussian mechanism is defined as:

Definition 2 (Gaussian Mechanism [53]): Let Δf be the sensitivity of function f, defined as $\Delta f = \max_{D,D'} \left\| f(D) - f(D') \right\|_2$. The Gaussian Mechanism, $\hat{f}(D) = f(D) + \sigma \Delta f \cdot \mathcal{N}(0, \mathcal{I})$, is deemed (ε, δ)-differentially private for specific values of ε and δ contingent on σ.

During our teacher models' training, random noise is introduced to perturb the original data distribution, thereby enhancing data privacy. Leveraging the post-processing property of differential privacy, as elucidated in [53], ensures that any subsequent operation on a differentially private output remains privacy-preserving. Thus, data generation under the guidance of the pre-trained teacher model is deemed secure. Specifically, random Gaussian noise is incorporated during the teacher model's training as follows:

$$g_T \leftarrow g_T + N(0, \sigma_n^2 c_g^2 I) \tag{3}$$

Here, g_T represents the teacher's gradients, σ_n is the noise scale, and c_g signifies the gradient function's sensitivity. Subsequently, the teacher model's weight parameters are

updated and truncated within the range (-c, c) to optimize the model:

$$w \leftarrow clip(w + \alpha \cdot Adam(w, g_T), -c, c) \tag{4}$$

For practical implementation, we use Opacus [55], Facebook's specialized library for training PyTorch models with differential privacy.

C. Dual Training Protocols

To address the multifaceted needs of data owners concerning both privacy preservation and performance optimization, two distinct training protocols have been devised, each characterized by a unique security level: the white-box and black-box protocols. Fig.3 provides a comprehensive visualization of the SG-ZSL paradigm within the context of both protocols. Each protocol encompasses three core components: 1) the isolated secure data and teacher model located at the data owner's domain; 2) the student model and generator positioned within the AI service provider's domain; and 3) the information exchange channels. For the white-box protocol, the teacher's weights are utilized in computing the gradient for both the generator and student network training. In contrast, the black-box protocol relies solely on the teacher model to furnish softmax output as pseudo labels, thereby excluding it from backpropagation during the SG-ZSL paradigm optimization.

1) White-Box Protocol: Under the white-box protocol, the data owner's pre-trained teacher model provides both gradient and softmax output to guide the training of models in the AI service providers, as explained below:

a) Uploading generated data: Under the guidance of the teacher model, the generator produces class-specific features using noise vector **z** and class-level semantic embedding **a** (attributes or BERT model representations of class names [2]). The synthesized features are represented as $\tilde{x} = G(z|a; \theta_G)$, with the objective of generating superior synthetic data instead of real data to train the student model.

b) Gradient and softmax guidance: The teacher model, upon receiving $\tilde{x}$, processes it through the loss function:

$$\min_{\theta_G} \mathcal{L}(\tilde{x}, y; \theta_G) + \alpha \mathcal{R}(\tilde{x}), \tag{5}$$

where $\mathcal{L}(\cdot)$ signifies the cross-entropy loss for classification by the teacher model, and $\mathcal{R}(\cdot)$ denotes the regularization term during feature generation. This regularization, executed at the data owner's domain, aims to minimize the distribution discrepancy between real and generated features, ensuring the AI service provider remains oblivious to the real data.

c) Feedback downloading: A request is dispatched to the AI service provider to retrieve the gradient, the regularization of distribution divergence, and the softmax output.

d) Label verification: Using the softmax output, pseudo labels are computed, and misclassified samples are filtered using label verification as follows:

$$(\tilde{x}^*, y^*) \in \{(\tilde{x}, y) | y = argmax\, T(\tilde{x}; \theta_T^*), \\ \tilde{x} = G(z|a; \theta_G^*)\}, \tag{6}$$

where T represents the teacher model, and $\tilde{x}^*$ and y^* denote the filtered high-quality generated features and their corresponding class labels, respectively.

e) Training the student model: Then the student model's training is articulated as follows:

$$\min_{\theta_S} \left\| T^*(\tilde{x}^*; \theta_T^*) - S(\tilde{x}^*; \theta_S) \right\|_2^2 \tag{7}$$

In this protocol, the gradient is directly applied to the generated features, markedly augmenting the generator's efficacy. However, the gradient feedback is deemed mid-risk information, potentially revealing details of the teacher model, while the softmax output and regularization feedback are categorized as low-risk.

2) Black-Box Protocol: The black-box protocol, in contrast to its white-box counterpart, restricts the teacher model's guidance during the second step. Specifically, only the softmax output and regularization can be solicited from the teacher model, ensuring the teacher model's weights remain inaccessible and that it is not involved in the backpropagation optimization process, thereby reducing the risk of model leakage. The black-box protocol is elucidated step-by-step below:

a) Uploading generated data: Analogous to the white-box protocol, the generated features are represented as $\tilde{x} = G(z|a; \theta_G)$, which are subsequently transmitted to the data owner for processing.

b) Softmax guidance: Upon receipt of the generated features $\tilde{x}$, the data owner calculates the softmax output and divergence regularization. Thereafter, a request is initiated to relay the feedback to the AI service provider.

c) Label verification: The generated data is then evaluated to ensure the conditional class input matches the teacher's softmax output, with misaligned samples discarded.

d) End-to-end training: The generator G and student network S undergo end-to-end training as outlined in the objective function:

$$\min_{\theta_G, \theta_S} \left\| T^*(\tilde{x}; \theta_T^*) - S(\tilde{x}; \theta_S) \right\|_2^2 + \alpha \mathcal{R}(\tilde{x}), \tag{8}$$

The comprehensive training procedures for both protocols are delineated in Algorithm 1.

D. Absolute Zero-Shot Classification

In the testing phase, the omniscient teacher, having been trained on both seen and unseen features at the Data Owner's end, facilitates the generator in synthesizing features for all classes. Consequently, the student network is equipped to predict class labels for test features. Given these test features, the predicted class labels are determined as:

$$y^* = \underset{y \in \mathcal{Y}}{argmax}\ p(y|x, \theta_S^*), \tag{9}$$

where θ_S^* represents the optimized parameters of the student model.

For the quasi-omniscient teacher model, the challenge confronting the student model intensifies. This heightened challenge arises because, during the training phase, neither the data owner nor the AI service provider possesses information

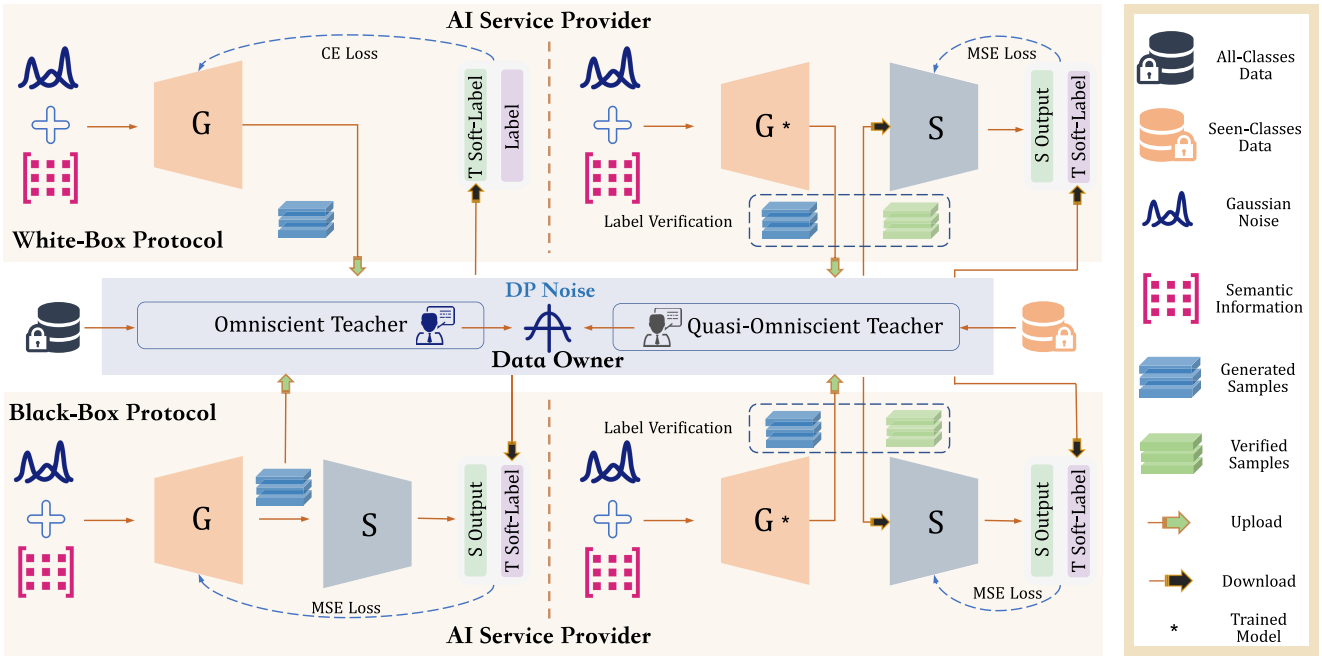


Fig. 3. The overarching paradigm for both black-box and white-box protocols. In the white-box protocol, the generator accesses teacher weights during training, whereas in the black-box protocol, only output guidance from the teacher is utilized.

regarding the unseen classes. In the testing phase, an initial step involves synthesizing a data batch for these unseen classes via the generator, denoted as $\tilde{x} = G(z|a;\theta_G^*)$, with z indicating noise and a representing the semantic embedding of the unseen class. Utilizing this synthesized data, the classifier C undergoes training in a supervised learning task with the generated features, as formalized in the following equation:

$$\min_{\theta_C} -\mathbb{E}\left[\log(y|\tilde{x};\theta_C))\right], \tag{10}$$

the function calculates the softmax loss by comparing the predicted label probabilities from synthesized features $\tilde{x}$ against actual labels y to minimize the negative log-likelihood of correct class predictions, optimizing classifier C for accurate unseen class label prediction.

Subsequently, the prediction of class labels for test features is executed as follows:

$$y^* = \underset{y\in\tilde{\mathcal{Y}}}{argmax}\ p(y|x,\theta_C^*), \tag{11}$$

where $\tilde{\mathcal{Y}} = \mathcal{Y}_u$ is designated for the conventional ZSL task, and $\tilde{\mathcal{Y}} = \mathcal{Y}_s \cup \mathcal{Y}_u$ for the GZSL task.

In the context of the first SG-ZSL work, this work primarily seeks to address the ensuing research questions:

- **RQ1:** How does the variation in teacher feedback influence the quality and diversity of the synthesized data?
- **RQ2:** How does the alteration in semantic information, when employed as generative conditions, affect the student model's performance?
- **RQ3:** Compare with the traditional ZSL methods, how do the SG-ZSL perform under the black-box and white-box protocols in terms of data privacy, model security, and classification accuracy?
- **RQ4:** Is the student model capable of transcending the constraints of the quasi-omniscient teacher model to generate novel knowledge (on unseen class)?
- **RQ5:** Does the SG-ZSL paradigm, which trains on both seen and unseen classes using synthesized data, enhance the congruence between seen and unseen classifiers in the GZSL challenge? Specifically, is there an improvement over prior ZSL approaches that employed real seen data and synthesized unseen data, potentially introducing a bias towards seen classes?

IV. Experiments

A. Datasets

Our SG-ZSL model is evaluated on three benchmark datasets: AWA1 [56], AWA2 [57], and aPY [58]. Both AWA1 and AWA2 encompass 30,475 and 37,322 images, respectively, distributed across 50 classes. The aPY dataset contains 15,539 images spanning 32 classes. For semantic representation, embeddings generated by the BERT language model [2] are adopted, with a consistent dimensionality of 768 across all datasets. The data splits differ based on the type of teacher model. For quasi-omniscient teachers, we adopt the data split proposed in [57], wherein only seen class data is accessible to the teacher. Conversely, the omniscient teacher is trained across all classes. In alignment with prior ZSL studies [59], unseen classes are randomly divided into training and test sets. Comprehensive dataset details and SG-ZSL data splits are presented in Table II.

B. Implementation Details

For image representation, 2048-dimensional ResNet101 features [1] are utilized, consistent with [57]. Within our proposed paradigm, all networks are constructed using Multi-Layer Perceptrons equipped with LeakyReLU activations [60].

TABLE II
DETAILED DATASET STATISTICS AND DATA SPLIT IN SG-ZSL. NOTATION: 'ATT' - ATTRIBUTE; 'S' - SEEN CLASS; 'U' - UNSEEN CLASS; 'OM' - OMNISCIENT TEACHER; 'Q-OM' - QUASI-OMNISCIENT TEACHER

Dataset	Semantics	Class Number	Image	Teacher (Om/Q-Om)		SG-ZSL Training		SG-ZSL Evaluation (Om/Q-Om)	
		S/U		S	U	S	U	S	U
AWA1 [56]	BERT/att	40/10	30475	19832	4542/0	**0**	**0**	4958	1143/5685
AWA2 [57]	BERT/att	40/10	37322	23527	6328/0	**0**	**0**	5882	1585/7913
aPY [58]	BERT/att	20/12	15539	5932	6333/0	**0**	**0**	1483	1591/7924

Algorithm 1 Training Procedure in Both Protocols

Require:
Pre-trained Teacher network θ_T^*, class labels $\mathcal{Y}_{tr}$ and their auxiliary semantic embedding $\mathcal{A}$; the maximal number of training epochs T_g and T_s for generator and student network, respectively.

Ensure:
The learned parameters θ_G, θ_S for generator G and student network S, respectively.

Initializing, θ_G, θ_S. Set the iteration epoch $t_g = 1$, $t_s = 1$.
while $t_g < T_g$ **do**
  if White-Box Protocol **then**
    Training generator with gradient guidance of teacher network through Eq.(5).
  else if Black-Box Protocol **then**
    Training generator with output guidance of teacher network through Eq.(8).
  end if
  $t_g := t_g + 1$;
end while
Conducting label verification through Eq.(6).
while $t_s < T_s$ **do**
  Training student network with output guidance of teacher through Eq.(7).
  $t_s := t_s + 1$;
end while

Both the teacher and student models share the same architecture comprising two hidden layers with 1024 and 512 units, respectively. The generator contains a single hidden layer with 4096 hidden units and its output layer is ReLU. During the training process, we adopt the Adam optimizer and the learning rate of each network is set to 10^{-5}. The dimension of the noise vector z is a hyper-parameter, which is empirically set to 20 for all datasets. The weight of the regularization term is empirically set to 0.5 for AWA1 and AWA2, and 1 for aPY. A trade-off between accuracy and computational efficiency is taken into consideration when determining the number of generated features. In practice, we generate 400 synthetic features on average per class for all datasets. For the training epochs T_g and T_s, we selected values that balance convergence and prevent overfitting or underfitting for both the generator and student network. Experimentally, we found performance plateaus in both networks beyond certain iterations, indicating an optimal stopping point for training. Consequently, $T_g = 50$ and $T_s = 80$ were set to optimize both computational efficiency and model effectiveness.

C. Evaluation Protocol

We follow the evaluation metrics proposed in [57]. For conventional ZSL tasks, we use the per-class average top-1 accuracy to evaluate classification performance to alleviate the data imbalance of classes. For the GZSL task, we use harmonic mean $H = (2 \times u \times s)/(u + s)$ for evaluation, where u and s denote average per-class top-1 accuracy on unseen and seen classes, respectively. It is noteworthy that existing methods aim to classify unseen data into corresponding unseen classes in conventional ZSL tasks, while the class space at test time involves both unseen and seen classes in SG-ZSL with the omniscient teacher. This makes SG-ZSL with an omniscient teacher more difficult compared with existing ZSL methods.

D. Main Results

1) Comparisons With State-of-Arts: Table III presents results for both CZSL and GZSL tasks. Given that this is the inaugural SG-ZSL study, a comparison with traditional state-of-the-art methods serves as a reference. The selected methods can be categorized into inductive and transductive ZSL methods. Methods in the upper part of Table III, *i.e.*, IAP, are inductive ZSL methods, which access only labeled seen class data during the training process. The rest of the four methods, *i.e.*, DTN, are transductive methods, which utilize both labeled seen class data and unlabeled unseen class data for model training.

To investigate **RQ1**, we show results under two kinds of feedback from omniscient and quasi-omniscient teachers. SG-ZSL student model with omniscient teacher achieves promising performance in both CZSL and GZSL in the white-box protocol. We achieve the best performance in GZSL, especially on aPY, with an increase in harmonic mean of 32.3%, which indicates an improved balance of seen and unseen classes. As for the black-box protocol, the accuracy on unseen classes is 4.9% higher than on seen classes on AWA1. It indicates that the SG-ZSL student model is promising to mitigate the class-level overfitting issue in the GZSL task proposed in **RQ5**. Compared with inductive ZSL methods, results show that our model with the quasi-omniscient teacher in a white-box protocol gains satisfactory performance in GZSL, especially on aPY, with 7.3% higher performance on the harmonic mean compared with non-generative inductive ZSL methods. Despite the quasi-omniscient teacher model's inability to recognize unseen classes and the student model's lack of access to real seen and unseen data, our student model still secures robust accuracy across various ZSL scenarios. For example, it achieves 34.5% accuracy in inductive ZSL settings on AWA1 and a harmonic mean of 26.7% in GZSL on the

TABLE III

COMPARISON RESULTS WITH THE STATE-OF-THE-ART METHODS IN CZSL AND GZSL TASKS. CZSL MEASURES PER-CLASS AVERAGE TOP-1 ACCURACY (T1) ON UNSEEN CLASSES. GZSL MEASURES U = T1 ON UNSEEN CLASSES, S = T1 ON SEEN CLASSES, H = HARMONIC MEAN. 'WB' & 'BB': WHITE- & BLACK-BOX PROTOCOL; 'OM' - OMNISCIENT TEACHER, 'Q-OM' - QUASI-OMNISCIENT TEACHER. 'SG-ZSL+WB/BB*' AND 'SG-ZSL+WB/BB' REPRESENT OUR MODEL WITH OMNISCIENT AND QUASI-OMNISCIENT TEACHERS, RESPECTIVELY. THE BEST RESULTS ARE IN BOLD

Method	Zero-Shot Learning			Generalized Zero-Shot Learning								
	AWA1	AWA2	aPY	AWA1			AWA2			aPY		
	T1	T1	T1	u	s	H	u	s	H	u	s	H
IAP [56]	35.9	35.9	36.6	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
DAP [56]	44.1	46.1	33.8	0.0	88.7	0.0	0.0	84.7	0.0	4.8	78.3	9.0
ALE [39]	59.9	62.5	39.7	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE [61]	54.2	59.7	39.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
CONSE [48]	45.6	44.5	26.9	0.4	88.6	0.8	0.5	90.6	1.0	0.0	**91.2**	0.0
ESZSL [45]	58.2	58.6	38.3	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC [62]	54.0	46.6	23.9	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
DEM [63]	68.4	67.1	35.0	32.8	84.7	47.3	30.5	86.4	45.1	11.1	75.1	19.4
f-CLSWGAN [38]	68.2	-	-	57.9	61.4	59.6	-	-	-	-	-	-
CE-GZSL [64]	71.0	70.4	-	65.3	73.4	69.1	63.1	78.6	70.0	-	-	-
SDGZSL [65]	-	74.3	47.0	-	-	-	69.6	78.2	73.7	39.1	60.7	47.5
ICCE [66]	74.2	72.7	49.5	67.4	81.2	73.6	65.3	82.3	72.8	45.2	46.3	45.7
DTN [59]	69.0	-	41.5	54.8	88.5	67.7	-	-	-	37.4	87.9	52.5
GMSADE [67]	81.3	80.7	49.9	71.2	87.7	78.6	71.3	86.1	78.0	76.1	39.3	51.8
EDE [68]	**85.3**	77.5	31.3	71.4	**90.1**	79.7	68.4	**93.2**	78.9	29.8	79.4	43.3
BGT [69]	-	**82.4**	49.8	-	-	-	56.2	82.2	66.7	39.3	72.9	51.0
Q-Om Teacher	0.0	0.0	0.0	0.0	92.9	0.0	0.0	93.1	0.0	0.0	91.6	0.0
Om Teacher	92.1	91.7	90.8	92.1	92.5	92.3	91.7	92.2	91.9	90.8	91.4	91.1
SG-ZSL+BB	14.1	19.9	12.3	4.1	3.7	3.9	3.5	3.7	3.6	6.8	4.0	5.1
SG-ZSL+WB	34.5	36.5	18.7	23.4	34.3	27.8	27.3	44.3	33.7	17.9	52.5	26.7
SG-ZSL+BB*	33.5	29.0	30.2	33.5	28.6	30.9	29.0	25.3	27.0	30.2	42.2	35.2
SG-ZSL+WB*	77.9	79.0	**83.9**	**77.9**	81.8	**79.8**	**79.0**	86.7	**82.7**	**83.9**	85.7	**84.8**

aPY dataset. This underscores the student model's capacity to extrapolate and generalize from the teacher's knowledge without data exposure, as explored in **RQ4**. Additionally, when contrasted with traditional TZSL methods, our model exhibits significant accuracy enhancements in GZSL, especially for unseen classes (*i.e.*demonstrate a 6.5% and 7.8% improvement on AWA1 and aPY datasets, respectively), and presents a reduced discrepancy between seen and unseen class accuracies, showcasing an advanced ability to mitigate seen class bias as mentioned in RQ5. For the black-box protocol, results show our SG-ZSL student model outperforms random guessing, which is around 10% on AWA1, AWA2, and 8% on aPY. The white-box protocol demonstrates better performance than the black-box protocol for the student, indicating that gradient guidance provides more information.

2) Comparisons in Black-Box Protocol: As it is the first time to propose this setting, we provide several baselines for comparison in Table IV. We provide labels and attributes for conditional feature generation to investigate **RQ2**. Our proposed paradigm with BERT embedding achieves the best performance, *i.e.*, with 18.0% and 23.4% increases in unseen accuracies on AWA1 compare with label-conditioned and attribute-contribution separately. Results show that our paradigm gains noticeable improvement in accuracy with label verification, *i.e.*, with 20.2% higher performance on harmonic mean on aPY dataset. And results indicate the effectiveness of adopting regularization, *i.e.*, it achieves 3.6% and 10.9% increases in Harmonic mean on AWA2 and aPY. The comparison with baselines demonstrates the effectiveness of our SG-ZSL student model in a black-box protocol with the omniscient teacher.

3) Performance vs Paradigm Privacy: Compared to traditional ZSL methods, the performance under the white-box protocol is very promising, since data privacy is already preserved and our model can still achieve adequate performance. Compare with the white-box protocol, the black-box protocol indeed operates under a more constrained information flow, where only softmax outputs from the teacher model are used as pseudo-labels for the student model, without direct gradient exchange. This design choice inherently poses challenges to optimization efficiency compared to direct gradient-based methods. However, this constraint is a deliberate design choice to enhance privacy. Thus, the performance of the black-box protocol is reasonable because both data privacy and model safety are guaranteed as proposed in **RQ3**.

As for model copyright reservation, traditional ZSL methods often involve sharing model details across entities, raising potential issues related to intellectual property and copyright infringement. Our SG-ZSL paradigm circumvents these issues by utilizing a sentinel mechanism that facilitates the learning process without exposing the internal architecture of the models involved. This is achieved by guiding the generation of synthetic data as a medium for communication between the AI Service provider and the Data Owner, enabling both parties without directly sharing the models themselves. This approach ensures that copyright and intellectual property rights are respected and protected, offering a sustainable model for collaborative AI development and usage.

E. Analysis and Discussion

1) Feature Generation Regularization Analysis: The key issue in our data-free knowledge transfer framework is to generate high-quality features, which are expected to have a similar distribution to real data. To show the influence of different constraints during the feature generation process, we provide analysis with different regularization terms for

TABLE IV
EXPERIMENTAL RESULTS IN BLACK-BOX PROTOCOL WITH THE OMNISCIENT TEACHER IN BOTH CZSL AND GZSL TASKS

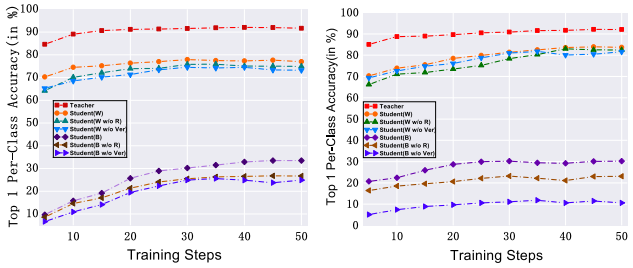
Method	Zero-Shot Learning			Generalized Zero-Shot Learning								
	AWA1	AWA2	aPY	AWA1			AWA2			aPY		
	T1	T1	T1	u	s	H	u	s	H	u	s	H
Label-Conditioned	15.5	10.0	7.0	15.5	24.3	18.9	10.0	17.8	12.8	7.0	3.8	4.9
Attribute-Conditioned	10.1	23.0	8.2	10.1	11.3	10.7	23.0	17.6	20.0	8.2	5.0	6.3
w/o Label Verification	25.6	24.7	11.8	25.6	15.6	19.4	24.7	18.1	20.9	11.8	20.9	15.0
w/o Regularization	26.8	23.7	23.2	26.8	26.7	26.8	23.7	23.2	23.4	23.2	25.6	24.3
SG-ZSL+BB	**33.5**	**29.0**	**30.2**	**33.5**	**28.6**	**30.9**	**29.0**	**25.3**	**27.0**	**30.2**	**42.2**	**35.2**

TABLE V
EXPERIMENTAL RESULTS WITH DIFFERENT CONSTRAINTS FOR FEATURE GENERATION IN GZSL TASK IN THE **WHITE-BOX** PROTOCOL. 'CE' REPRESENTS CROSS-ENTROPY LOSS, 'MMD' REPRESENTS MMD DISTANCE LOSS, AND 'KL' REPRESENTS KL DIVERGENCE LOSS

Method	AWA2			aPY		
	u	s	H	u	s	H
CE	76.1	83.8	79.8	83.0	84.5	83.7
CE+MMD	**79.9**	85.1	82.5	81.5	85.5	83.5
CE+KL	79.0	**86.7**	**82.7**	**83.9**	**85.7**	**84.8**

TABLE VI
EXPERIMENTAL RESULTS WITH DIFFERENT CONSTRAINTS FOR FEATURE GENERATION IN GZSL TASK IN THE **BLACK-BOX** PROTOCOL. 'CE' REPRESENTS CROSS-ENTROPY LOSS, 'MMD' REPRESENTS MMD DISTANCE LOSS, AND 'KL' REPRESENTS KL DIVERGENCE LOSS

Method	AWA1			AWA2			aPY		
	u	s	H	u	s	H	u	s	H
CE	26.8	26.7	26.8	23.7	23.2	23.4	23.2	25.6	24.3
CE+MMD	31.8	25.3	28.2	**33.8**	20.5	25.5	26.0	36.3	30.3
CE+KL	**33.5**	**28.6**	**30.9**	29.0	**25.3**	**27.0**	**30.2**	**42.2**	**35.2**



(a) AWA1 (b) aPY

Fig. 4. Epoch analysis for unseen accuracy. 'Ver': label verification. 'R': regularization term.

TABLE VII
RESULTS IN THE WHITE-BOX PROTOCOL WITH AN OMNISCIENT TEACHER UNDER DIFFERENT PRIVACY BUDGETS ϵ

Dataset	Accuracy	$\epsilon = 30$	$\epsilon = 50$	$\epsilon = \infty$
AWA1	Teacher Model	56.7	68.4	92.1
	Harmonic Mean	41.7	56.4	79.8
AWA2	Teacher Model	59.1	70.5	91.7
	Harmonic Mean	46.8	60.3	82.7
aPY	Teacher Model	60.6	72.4	90.8
	Harmonic Mean	43.6	62.2	84.8

generator training in Table V. KL and MMD loss [70] aim to minimize the distribution difference between real and generated features. Results show that adding distribution constraints of synthesized data is beneficial for feature generation. For example, the harmonic mean increases 2.7% and 2.9% with MMD and KL loss respectively compared with the baseline that only contains cross-entropy loss. Besides, results indicate that KL and MMD loss are both effective and KL loss performs better to a small extent, which shows the effectiveness of KL regularization.

We also provide an extensive analysis of the impact of different feature generation regularizations in the black-box scenario in Table VI. Similarly, we provide MMD and KL loss as regularization for feature synthesis in the GZSL task as the regularization term is essential for the generalization ability of the SG-ZSL model. The experimental results show that the SG-ZSL model with regularization term outperforms the one with only cross-entropy loss, *i.e.*, with 6% and 10.9% improvement on harmonic mean with MMD and KL loss on aPY, indicating the effectiveness of the constraint for feature generation. Besides, the SG-ZSL model with KL constraint achieves the best performance in harmonic mean, with 4.9% and 2.7% increases on aPY and AWA1 datasets respectively, which indicates that the SG-ZSL model with KL loss can make a better balance between seen and unseen classes.

2) Student vs Teacher Performance Analysis: Here, we undertake experiments to assess the relationship between teacher performance and student performance. The outcomes observed from the omniscient teacher and student models, across escalating training steps in both protocols on AWA1 and aPY, are depicted in Fig. 4.

Operating under supervised learning with access to real data, our teacher model reliably guides student models, as evidenced by its performances of 92.1% on the AWA1 dataset and 90.8% on the aPY dataset. Within the white-box protocol, the student model approaches the performance of the teacher model, signifying the efficacy of the gradient guidance mechanism. Furthermore, our findings elucidate that the incorporation of regularization terms enhances the model's performance, showcasing the pivotal role of feature distribution throughout the training phase. Further, our analysis reveals that label verification in both protocols enhances performance, highlighting its necessity. This is attributed to its capacity to mitigate the negative effects resulting from the creation of low-quality features.

3) Teacher Model Privacy Evaluation: Table VII displays the performance corresponding to various privacy budgets ϵ when DP is incorporated into teacher training. Here, $\epsilon = \infty$ signifies the baseline non-private performance, i.e., absent DP in teacher training. The results demonstrate that larger ϵ values

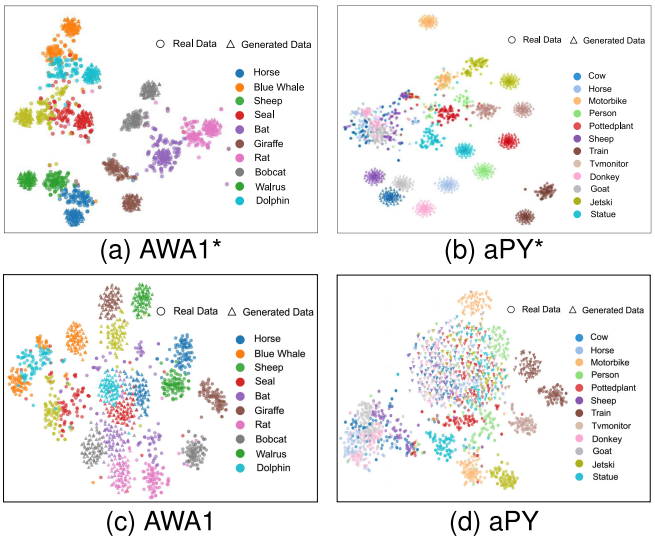


Fig. 5. The t-SNE visualization on AWA1 and aPY. All experiments are simulated under white-box protocol, with the synthetic features in (a) and (b) generated from generators that follow the omniscient teacher (indicated with *), and those in (c) and (d) generated from generators that follow the quasi-omniscient teacher.

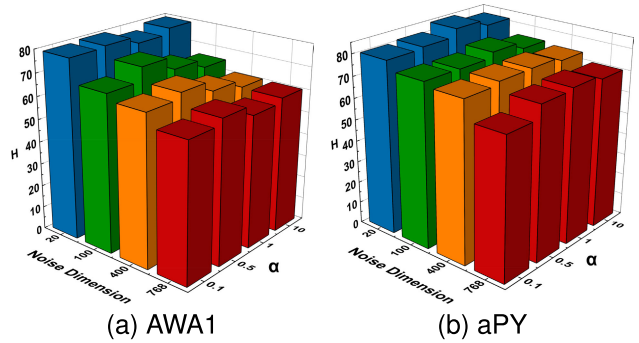


Fig. 6. Noise dimension and parameter α analysis with omniscient teacher in white-box protocol.

correspond to enhanced performances for both the teacher and student models, indicating that a smaller ϵ yields heightened data security protection. A trade-off between performance and privacy level is observed, allowing for an adjustment of the privacy budget to achieve a balance.

4) Quality of Generated Features: Fig. 5 displays t-SNE visualizations of real and synthetic unseen features under the white-box protocol guided by two distinct teacher models across AWA1 and aPY datasets. For clarity in visualization, a subset of features is randomly selected. Features synthesized under the guidance of the omniscient teacher, illustrated in (a) and (b), closely emulate real feature distributions and demonstrate significant class clustering. This effectiveness, even without real data, highlights our model's capacity to generate class-coherent features in ZSL scenarios. In comparison, unseen class features synthesized under the guidance of the quasi-omniscient teacher, shown in (c) and (d), exhibit a slight decline in quality, *i.e.*, the distribution of generated features is farther from the real data distribution, which illustrates the limitation of the generator. However, the model is trained without access to unseen class information. The capability of synthesizing unseen class data shows the potential to generate novel knowledge.

TABLE VIII

EXPERIMENTAL RESULTS IN WHITE-BOX PROTOCOL WITH OMNISCIENT TEACHER USING DIFFERENT SEMANTIC INFORMATION IN GZSL TASK

Semantics	AWA1			AWA2		
	u	s	H	u	s	H
Attribute	64.7	81.1	72.0	76.8	82.7	79.6
Word2vec	61.6	80.0	69.5	71.4	81.9	76.3
BERT	**77.9**	**81.8**	**79.8**	**79.0**	**86.7**	**82.7**

TABLE IX

RESULTS WITH DIFFERENT STUDENT MODELS IN BLACK-BOX PROTOCOL WITH OMNISCIENT TEACHER IN GZSL TASK

Student Model	**AWA1**			**aPY**		
	u	s	H	u	s	H
1 hidden layer	31.9	25.9	28.6	**30.4**	34.4	32.3
3 hidden layer	27.9	26.3	27.1	28.5	36.4	32.0
Ours	**33.5**	**28.6**	**30.9**	30.2	**42.2**	**35.2**

5) Hyper-Parameter Analysis: We assess the impact of two pivotal hyper-parameters, namely, noise dimension and regularization weight, on our student model. Two ablation studies are conducted on the AWA1 and aPY datasets within a white-box protocol framework, engaging an omniscient teacher, as illustrated in Fig. 6. We select four disparate noise dimensions 20, 100, 400, and 768 to elucidate their relationship with the harmonic mean. The findings reveal a performance decrement correlating with the expansion of the noise dimension across both datasets, suggesting that higher-dimensional noise may engender significant interference. Concerning the regularization weight, we designate the values of α as 0.1, 0.5, 1, and 10 for the experimental analysis. As shown in Fig. 6, the harmonic mean on both datasets exhibits marginal fluctuation with varying α values. Optimal performance is attained at α values of 0.5 and 1 for AWA1 and aPY datasets, respectively, demonstrating a nuanced interaction between regularization weight and model performance.

6) Impact of Semantic Information: We further investigate the influence of various semantic embeddings on the GZSL task. The experimental analysis encompasses three distinct semantic typologies, namely, attributes, Word2Vec, and BERT, serving as the evaluation benchmarks. As delineated in Table VIII, the comparative outcomes across all three semantic modalities in the GZSL task are relatively aligned, manifesting the robustness of our model with respect to semantic embedding. Notably, the BERT embedding outperforms, signifying the superior efficacy of BERT representation in capturing semantic nuances.

7) Robustness of Student Network: We elucidate the robustness inherent to the student network in this section. Given that the teacher network remains undisclosed by the Data Owner within the black-box protocol, it becomes imperative to showcase the results across diverse student models in this black-box scenario. As illustrated in Table IX, the performances across various student models are closely aligned, denoting the stability and consistency afforded by our method.

F. Potential Applications

As for potential applications, our SG-ZSL paradigm could carry profound implications for industries where data privacy is paramount. In healthcare, SG-ZSL can facilitate the sharing of medical insights without exposing patient data, thus advancing research while complying with stringent confidentiality regulations. Similarly, in finance, SG-ZSL enables the collaborative development of predictive models without risking sensitive financial information. Consequently, SG-ZSL fosters a collaborative environment where both data owners and AI service providers can thrive, leveraging the strengths of each party without compromising on security or copyright.

G. Limitations

Although our research raises awareness of data and model privacy in the ZSL field, balancing privacy with performance remains challenging. The white-box protocol offers high performance through the guidance of teacher model weights and outputs but demands a careful balance between privacy and performance using differential privacy techniques. Meanwhile, the inherently secure black-box protocol may lag in optimization and performance due to its exclusive reliance on output-based supervision. Future efforts aim to bridge these gaps by enhancing the generator's capabilities, notably by incorporating common-sense knowledge from large-scale models to establish a more robust knowledge space, thus improving knowledge transfer from seen to unseen classes. Additionally, our proposed SG-ZSL framework presently lacks a comprehensive assessment of computational efficiency and time costs. We aim to tackle this in our forthcoming work by meticulously quantifying these critical metrics and honing in on minimizing communication expenses. Enhancements will also focus on improving the visualization of generated data, further refining our approach to seamlessly integrate privacy concerns with optimal performance.

V. Conclusion

In this work, we introduced an SG-ZSL paradigm facilitating through data-free knowledge transfer. A pre-trained teacher model was instantiated at the data owner's end, acting as a data sentinel to render guidance for model training. A thorough evaluation was conducted for both 'black-box' and 'white-box' protocols, elucidating the trade-off between model performance and data privacy. Based on the proposed paradigm, the real data does not participate in the training at the AI service provider end, our model exhibits comparable performance against CZSL and GZSL while the data privacy is also secured. Future advancements in SG-ZSL can explore advanced optimization strategies based on more representative common knowledge (*i.e.*from Large Language Models), and investigate more robust privacy protections, ensuring data owner interests are preserved without compromising model performance.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[3] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.

[4] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput. (TAMC)*. Cham, Switzerland: Springer, 2008, pp. 1–19.

[5] L. Zhang, G. Gao, and H. Zhang, "Spatial–temporal federated learning for lifelong person re-identification on distributed edges," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[6] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2021–2031.

[7] F. Wan, J. Wang, H. Duan, Y. Song, M. Pagnucco, and Y. Long, "Community-aware federated video summarization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023, pp. 1–8.

[8] H. Ren, J. Deng, X. Xie, X. Ma, and Y. Wang, "FedBoosting: Federated learning with gradient protected boosting for text recognition," *Neurocomputing*, vol. 569, Feb. 2024, Art. no. 127126.

[9] S. Wang et al., "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[10] R. Yu and P. Li, "Toward resource-efficient federated learning in mobile edge computing," *IEEE Netw.*, vol. 35, no. 1, pp. 148–155, Jan. 2021.

[11] S. Guo, T. Zhang, G. Xu, H. Yu, T. Xiang, and Y. Liu, "Topology-aware differential privacy for decentralized image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 4016–4027, Jun. 2022.

[12] N. Fu, W. Ni, S. Zhang, L. Hou, and D. Zhang, "GC-NLDP: A graph clustering algorithm with local differential privacy," *Comput. Secur.*, vol. 124, Jan. 2023, Art. no. 102967.

[13] G. Xu, G. Li, S. Guo, T. Zhang, and H. Li, "Secure decentralized image classification with multiparty homomorphic encryption," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3185–3198, Jul. 2023.

[14] G. Xu, Z. Liu, and C. Change Loy, "Computation-efficient knowledge distillation via uncertainty-aware mixup," 2020, *arXiv:2012.09413*.

[15] H. Liu, X. Zhu, Z. Lei, D. Cao, and S. Z. Li, "Fast adapting without forgetting for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3093–3104, Aug. 2021.

[16] K. Xu, L. Wang, J. Xin, S. Li, and B. Yin, "Learning from teacher's failure: A reflective learning paradigm for knowledge distillation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 384–396, 2024.

[17] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10915–10924.

[18] K. Zhang, C. Zhang, S. Li, D. Zeng, and S. Ge, "Student network learning via evolutionary knowledge distillation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2251–2263, Apr. 2022.

[19] H. Zhang, Y. Long, Y. Guan, and L. Shao, "Triple verification network for generalized zero-shot learning," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 506–517, Jan. 2019.

[20] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *Proc. AAAI*, vol. 1, 2008, p. 3.

[21] M. R. Vyas, H. Venkateswara, and S. Panchanathan, "Leveraging seen and unseen semantic relationships for generative zero-shot learning," in *Proc. ECCV*, 2020, pp. 70–86.

[22] R. Gao et al., "Privacy-enhanced zero-shot learning via data-free knowledge transfer," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 432–437.

[23] D. Cheng, G. Wang, B. Wang, Q. Zhang, J. Han, and D. Zhang, "Hybrid routing transformer for zero-shot learning," *Pattern Recognit.*, vol. 137, p. 109270, 2023.

[24] D. Cheng, G. Wang, N. Wang, D. Zhang, Q. Zhang, and X. Gao, "Discriminative and robust attribute alignment for zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4244–4256, 2023.

[25] P. Huang, J. Han, D. Cheng, and D. Zhang, "Robust region feature synthesizer for zero-shot object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7622–7631.
[26] D. Jayaraman and K. Grauman, "Zero-shot recognition with unreliable attributes," in *Proc. NeurIPS*, vol. 27, 2014.
[27] S. Li, L. Wang, S. Wang, D. Kong, and B. Yin, "Hierarchical coupled discriminative dictionary learning for zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4973–4984, 2023.
[28] Y. Long, L. Liu, and L. Shao, "Towards fine-grained open zero-shot learning: Inferring unseen visual features from attributes," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 944–952.
[29] Y. Guo, G. Ding, J. Han, and S. Tang, "Zero-shot learning with attribute selection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018.
[30] Z. Zhang and V. Saligrama, "Zero-shot recognition via structured prediction," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 533–548.
[31] Y. Long and L. Shao, "Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 907–915.
[32] H. Zhang, Y. Long, and L. Shao, "Zero-shot leaning and hashing with binary visual similes," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 24147–24165, Sep. 2019.
[33] J. Qin, Y. Wang, L. Liu, J. Chen, and L. Shao, "Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1667–1671, Nov. 2016.
[34] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4447–4456.
[35] R. Felix, V. B. Kumar, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. ECCV*, 2018.
[36] J. Wang, Y. Jiang, Y. Long, X. Sun, M. Pagnucco, and Y. Song, "Deconfounding causal inference for zero-shot action recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 21–37, 2024.
[37] R. Gao et al., "Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3665–3680, 2020.
[38] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5542–5551.
[39] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 819–826.
[40] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2927–2936.
[41] Y. Tian, Y. Kong, Q. Ruan, G. An, and Y. Fu, "Aligned dynamic-preserving embedding for zero-shot action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1597–1612, Jun. 2020.
[42] Y. Liu, Q. Gao, J. Li, J. Han, and L. Shao, "Zero shot learning via low-rank embedded semantic AutoEncoder," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, p. 10.
[43] Y. Liu, X. Gao, J. Han, L. Liu, and L. Shao, "Zero-shot learning via a specific rank-controlled semantic autoencoder," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108237.
[44] Y. Long, L. Liu, L. Shao, F. Shen, G. Ding, and J. Han, "From zero-shot learning to conventional supervised classification: Unseen visual data synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6165–6174.
[45] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. ICML*, 2015, pp. 2152–2161.
[46] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, "Transductive unbiased embedding for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1024–1033.
[47] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332–2345, Nov. 2015.
[48] M. Norouzi et al., "Zero-shot learning by convex combination of semantic embeddings," in *Proc. ICLR*, 2014.
[49] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proc. ECCV*, 2016, pp. 52–68.
[50] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12661–12670.
[51] Y. Hu, L. Feng, H. Jiang, M. Liu, and B. Yin, "Domain-aware prototype network for generalized zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.
[52] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Automatic clipping: Differentially private deep learning made easier and stronger," 2022, *arXiv:2206.07136*.
[53] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
[54] L. Zhang, B. Shen, A. Barnawi, S. Xi, N. Kumar, and Y. Wu, "FedDPGAN: Federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia," *Inf. Syst. Frontiers*, vol. 23, no. 6, pp. 1403–1415, Dec. 2021.
[55] A. Yousefpour et al., "Opacus: User-friendly differential privacy library in PyTorch," 2021, *arXiv:2109.12298*.
[56] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.
[57] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," in *Proc. CVPR*, 2017, pp. 2251–2265.
[58] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1778–1785.
[59] H. Zhang, L. Liu, Y. Long, Z. Zhang, and L. Shao, "Deep transductive network for generalized zero shot learning," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107370.
[60] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," in *Proc. ICML Deep Learn. Workshop*, 2015.
[61] A. Frome et al., "Devise: A deep visual-semantic embedding model," in *Proc. NeurIPS*, vol. 22, 2013.
[62] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha, "Synthesized classifiers for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5327–5336.
[63] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3010–3019.
[64] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2371–2381.
[65] Z. Chen et al., "Semantics disentangling for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8692–8700.
[66] X. Kong et al., "En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9296–9305.
[67] O. Gune, M. Pal, P. Mukherjee, B. Banerjee, and S. Chaudhuri, "Generative model-driven structure aligning discriminative embeddings for transductive zero-shot learning," 2020, *arXiv:2005.04492*.
[68] L. Zhang et al., "Towards effective deep embedding for zero-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2843–2852, Sep. 2020.
[69] X. Li, D. Zhang, M. Ye, X. Li, Q. Dou, and Q. Lv, "Bidirectional generative transductive zero-shot learning," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 5313–5326, May 2021.
[70] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

Fan Wan received the M.Sc. degree (Hons.) in computer science from Newcastle University, U.K., in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Durham University. His research interests include federated learning, zero-shot learning, and video summarization.

Xingyu Miao received the master's degree from the School of Information Engineering, Ningxia University, China. He is currently pursuing the Ph.D. degree with the Department of Computing, Durham University, U.K. His current research interests include monocular depth estimation and 3D reconstruction.

Haoran Duan (Graduate Student Member, IEEE) received the M.S. degree in data science from Newcastle University, U.K., in 2019. He is currently pursuing the Ph.D. degree with Durham University, U.K. He was a research student with the Open Laboratory, Newcastle University. He is a part-time Research Associate with the Network and Ubiquitous Systems Engineering Group, School of Computing, Newcastle University, working on deep learning applications on edge. His current research interests include the applications/theories of deep learning.

Jingjing Deng received the Ph.D. degree in visual computing from Swansea University, U.K., in 2017. He is currently an Assistant Professor with the Vision, Imaging, and Visualization in Durham (VIViD) Group, Department of Computer Science, Durham University, U.K. His research interests include computer vision and machine learning. He and his team focus on developing computational models that can cultivate and generalize intelligence from and for the complex world. The team is actively seeking collaborative opportunities motivated by industrial practices and theoretical principles across disciplines.

Rui Gao received the B.S. degree from the School of Electronics and Information Engineering, Sichuan University, China, in 2015, and the Ph.D. degree from the School of Electronic and Information Engineering, Xi'an Jiaotong University, China, in 2023. Currently, she is a Research Associate with the School of Computing and Mathematic Sciences, University of Leicester, U.K. Her current research interests include zero-shot learning, multi-modal learning, and computer vision.

Yang Long (Senior Member, IEEE) is currently an Assistant Professor with the Department of Computer Science, Durham University. He is also an MRC Innovation Fellow aiming to design scalable AI solutions for large-scale healthcare applications. He has authored/coauthored more than 30 top-tier papers in refereed journals/conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, CVPR, AAAI, and ACM MM. He holds a patent and a Chinese National Grant. His research interests include computer vision and machine learning. He is passionate about unveiling the black box of the AI brain and transferring the knowledge to seek scalable, interactable, interpretable, and sustainable solutions for other disciplinary research, such as physical activity, mental health, design, education, security, and geoengineering.