2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00081

PRIVACY-ENHANCED ZERO-SHOT LEARNING VIA DATA-FREE KNOWLEDGE TRANSFER

Rui Gao[*,1], *Fan Wan*[*,3], *Daniel Organisciak*[2], *Jiyao Pu*[3], *Haoran Duan*[3], *Peng Zhang*[3], Xingsong Hou[1], Yang Long[*,3]

[1]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[2]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK
[3]Department of Computer Science, Durham University, Durham, UK

ABSTRACT

Considering the increasing concerns about data copyright and sensitivity issues, we present a novel Privacy-Enhanced Zero-Shot Learning (PE-ZSL) paradigm. The key innovation is to involve a teacher model as the data safeguard to guide the PE-ZSL model training without data sharing. The PE-ZSL model consists of a generator and student network, which can achieve data-free knowledge transfer while maintaining the performance of teacher model. We investigate 'black-' and 'white-box' scenarios in PE-ZSL task as different levels of framework privacy. Besides, we provide the discussion of teacher model in both omniscient and quasi-omniscient settings according to the knowledge space. Despite simple implementations and data-missing disadvantages, our PE-ZSL framework can retain state-of-the-art ZSL and GZSL performance under the 'white-box' scenario. Extensive qualitative and quantitative analysis also demonstrates promising results when deploying the model under 'black-box' scenario.

Index Terms— Zero-Shot Learning, Privacy Protection, Data-Free Knowledge Transfer

1. INTRODUCTION

The blossom of deep learning technologies embraces the development of high-performance computing and large-scale multi-modal data. However, sharing data across different institutes and even between different countries has become increasingly difficult and sensitive. The increasing awareness of data copyright, expensive data annotation, and restricted access to data in expert domains have hindered the development of interdisciplinary and intercultural deep models.

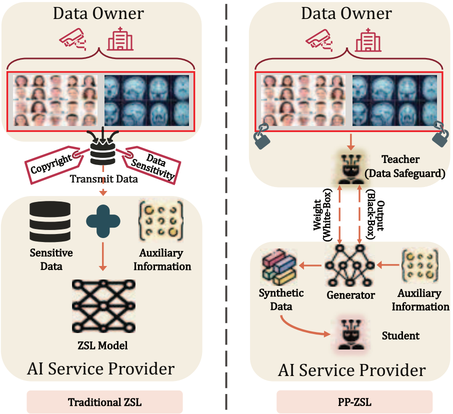


Fig. 1: Differences between ZSL and PE-ZSL. ZSL models require real data to learn visual-semantic association. PE-ZSL suggests a teacher model as data safeguard to tackle ZSL tasks without access to real data.

As shown in Fig.1, datasets may contain sensitive data, *i.e.*, healthcare and face information, which cost data owners over billions and tens of years to collect. Strict regulations, such as the GDPR [1] in Europe have been enforced to control the risk of data leaking. **In our paper, we focus on protecting data copyright and eliminating data sensitivity from data owners when data contain confidential user information.** Concretely, AI service providers (*i.e.*, AI companies) maintain close cooperation with the data owners (*i.e.*, scientific institutions and hospitals) and it needs to obtain data to provide related services for customers. However, the healthcare dataset is expensive to collect so the hospital cannot share the data directly with AI service providers due to copyright protection. When AI companies need access to such sensitive data to provide AI services, the shared data is exposed to leaking risks even though tedious confidentiality agreements have been signed. Take another example of the situation related to surveillance data, AI service providers take on the task, *i.e.*, pedestrian re-identification, while it is necessary to eliminate the sensitive user face information. **Motivated by these challenges, this paper explores a theoretical case when an AI developer needs to train a new model, the data owner (from different institutes) can provide a data-free teacher service as an API so that knowledge can be transferred**

[*]Contributed Equally.

Corresponding Author: Yang Long (yang.long@durham.ac.uk).

This project is supported by International Exchanges 2022 IEC\NSFC \223523 and Securing the Energy/Transport Interface EP/X037401/1, National Natural Science Foundation of China (62272376; 61872286), Key R&D Program of Shaanxi Province (2020ZDLGY04-05; S2021-YF-YBSF-0094).

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00081

without any data sharing.

As a promising machine learning paradigm, zero-shot learning (ZSL) shows the good potential to tackle data-free problems which investigates an extreme case when the such deep transfer can go beyond seen classes in the teacher dataset. Existing ZSL models are established based on real data from either seen or unseen classes. When adapting a pre-trained model to a new task domain, existing ZSL models assume a large amount of labeled seen class or unlabeled unseen class data are available to establish the visual-semantic relationship. However, sharing data across different institutes and even different countries is often infeasible. Different from existing ZSL settings, we focus on establishing ZSL model without data sharing during training. In this paper, we propose a new paradigm dubbed Privacy-Enhanced Zero-Shot Learning (PE-ZSL) to avoid sensitive data leaking while still enabling AI model can be trained. Figure 1 briefly illustrates the difference between ZSL and PE-ZSL task. Our PE-ZSL task suggests replacing data with a teacher model (pre-trained on real data) to guide the ZSL model training. Teacher model can be regarded as the implicit representation of data so that PE-ZSL model can be established through the supervision of teacher model, which can prevent real data from being shared.

To comprehensively explore our proposed PE-ZSL framework, we also present extensive discussion from the perspective of privacy issues and knowledge space of the teacher model. First, we propose two PE-ZSL scenarios in terms of framework privacy. In the 'black-box' scenario, teacher only provides output classification scores but does not share weights. In 'white-box' scenario, teacher will also share the model weights during training, which is more informative. These two scenarios indicate different levels of communication between data owners and AI service providers, which will lead to different ZSL recognition performances. In terms of teacher model privacy, we adopt differential privacy (DP) [2] in teacher training, which provides protection against the adversary who has access to model information, *i.e.*, parameters. Furthermore, we propose omniscient and quasi-omniscient teachers according to the knowledge space, *i.e.*, whether unseen classes are involved in pre-training teacher model. In summary, our contributions are three-fold:

- Privacy-Enhanced Zero-Shot Learning aims to achieve zero-shot classification without access to real data. The paradigm can be applied to real-world applications for data copyright protection and sensitivity elimination.
- We develop a novel data-free knowledge transfer framework for PE-ZSL task. In addition to zero data sharing setting, we propose 'black-' and 'white-box' scenarios and discuss the pros and cons of model sharing problems. We also present analysis of teacher model in both omniscient and quasi-omniscient settings according to the knowledge space.
- We show experimental results for conventional and generalized ZSL tasks in two scenarios. Though PE-ZSL model is established without data sharing during training, it achieves promising performance.

2. RELATED WORK

The most widely used framework for data privacy enhancing is Federated Learning [3]. A global model is shared with clients to avoid data leaking. Knowledge distillation [4] utilizes the domain-expert teacher model to train a compact student model and it can prevent teacher model from being attacked. Yet, none of these methods have explored potential in zero-shot learning situation. This paper presents the first work exploring privacy-enhanced zero-shot learning paradigm via data-free knowledge transfer. Besides, the differential privacy (DP) [2] technology can be applied for teacher model for data privacy protection further.

Zero-Shot learning [5] enables deep learning model[6, 7, 8, 9] to recognise unknown/unseen classes by establishing the relationship between seen and unseen classes via class semantic information. Some works [10] aim to build the mapping between visual and semantic space. Other works [11] focus on unseen class data generation to alleviate data-missing problem. According to whether unseen data is adopted during training, existing ZSL methods can be categorised into inductive [8] and transductive [12] settings. As for test phase, conventional ZSL (CZSL) methods [13] assume test data only come from unseen classes, while generalized ZSL (GZSL) [14] is then proposed to assign both seen and unseen data into corresponding classes. There has been little research on zero-shot learning whilst enhancing data privacy, so we propose a privacy-enhanced zero-shot learning paradigm, which aims to accomplish the zero-shot recognition without access to real data during training.

3. PRIVACY-ENHANCED ZERO-SHOT LEARNING

As shown in Fig.2, PE-ZSL addresses the problem when sensitive data is secured on the *Data Owner* domain. The key idea is to introduce a teacher model as the data safeguard and guide the model deployed on the *AI Service Provider* domain to train a classifier with zero real data. In addition to data privacy enhancing, we introduce white- and black-box scenarios to discuss the teacher model sharing problem regarding the balance between performance and security.

3.1. Problem Formulation

The basic PE-ZSL setting involves secured images and their extracted visual features $x \in \mathcal{X}$. The data safeguard is a pre-trained teacher model on the data owner domain. For simplicity, we consider supervised learning model $f_T : \mathcal{X} \rightarrow \mathcal{Y}$, where $y \in \mathcal{Y}$ is the label space. The ultimate goal is to train a student model on AI service provider domain using an objective function ℓ that learns from the guidance of teacher model:

$$\ell\left(f_{PE-ZSL}\left(\tilde{x}\right), f_T\left(\tilde{x}\right)\right), \tag{1}$$

where $\tilde{x} \in \tilde{\mathcal{X}}$ is generated data that ensures no real data can be accessed.

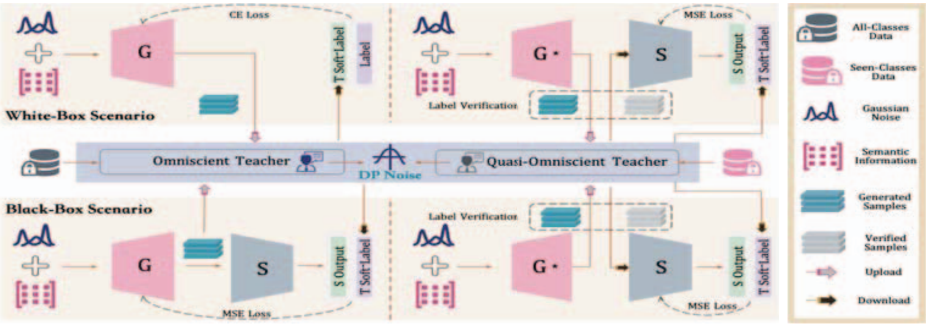


Fig. 2: Overall framework in the black-box and white-box scenarios. In white-box scenario, the generator has access to teacher weights during training while teacher only provides output guidance in black-box scenario.

PE-ZSL with Omniscient & Quasi-Omniscient Teacher On the data owner domain, we further breakdown the PE-ZSL into omniscient and quasi-omniscient teachers according to the label space. Seen classes are defined 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. Seen and unseen classes are disjoint, *i.e.*, $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$. For PE-ZSL, both seen and unseen features, x_s and x_u, are unavailable for service provider. Available information for service provider can be represented as $\mathcal{T}_r = \{(a, y) \mid a \in \mathcal{A}, y \in \mathcal{Y}\}$, which means only semantic embedding and class labels can be accessed. Teacher model pre-trained by real data is provided for model training guidance. In this way, the basic PE-ZSL with omniscient teacher considers $f_T : \mathcal{X} \to \mathcal{Y}$ because the source domain contains both seen and unseen classes. A more challenging PE-ZSL with quasi-omniscient teacher considers $f_T : \mathcal{X}_s \to \mathcal{Y}_s$.

ZSL vs GZSL On AI service provider domain, PE-ZSL aims to classify test images $f_{ZSL} : \mathcal{X}_u \to \mathcal{Y}_u$ for CZSL, and $f_{GZSL} : \mathcal{X} \to \mathcal{Y}$ for GZSL. Training of above classifiers using absolutely generated data will be introduced next.

3.2. White-Box & Black-Box Scenarios

The objective function of PE-ZSL in Eq.(1) defines a data-free knowledge transfer framework for PE-ZSL task, *i.e.*, through the guidance of the teacher model, our proposed PE-ZSL formula consists of two components: a *Generator* G and a *Student* network S. Generator G is to synthesize features and student S aims to match the performance of teacher model. Figure 2 depicts the detailed PE-ZSL framework in both white-box and black-box scenarios. The system consists of 1) the secured data and teacher model on the data owner; 2) PE-ZSL model on the AI service provider; 3) and the information exchange channels. Considering the model inversion can attack shared models, we also investigate the security levels regarding model sharing in addition to data privacy enhancing. For white-box scenario, the teacher weights are involved to compute the gradient for the generator and student network training. In the black-box scenario, the teacher only provides the output as pseudo labels, *i.e.*, teacher model is not involved in back propagation for optimization of PE-ZSL framework.

White-Box Scenario. In white-box scenario, the teacher model provides both gradient and softmax output as the PE-ZSL training guidance as follows. **1) Uploading generated data:** the generator synthesizes features of same classes with the supervision with teacher based on noise **z** and class-level semantic embedding **a** (attributes or BERT model of class names [15] as the condition). Specifically, we aim to synthesize the features that can be classified into corresponding classes with the constraint of the teacher network. $\tilde{x} = G(z|a; \theta_G)$ represents the generated features which are uploaded to the data owner. **2) Gradient and softmax guidance:** The teacher model receives $\tilde{x}$ and process the data using the loss function:

$$\min_{\theta_G} \mathcal{L}(\tilde{x}, y; \theta_G) + \alpha \mathcal{R}(\tilde{x}), \tag{2}$$

where $\mathcal{L}(\cdot)$ represents cross-entropy loss by teacher model for classification, $\mathcal{R}(\cdot)$ refers to the regularization term during feature generation with hyperparameter α. The regularization term aims to minimize the distribution distance of real and generated features. Note that regularization is also completed at the data owner side and real data will not be accessed for the AI service provider. **3) Feedback downloading**: a request is sent to service provider so that the gradient, the regularization of distribution divergence and softmax output can be download. **4) Label verification:** Using softmax to compute pseudo labels and filter out misclassified generated samples:

$$\begin{aligned}(\tilde{x}^*, y^*) \in \{(\tilde{x}, y) | y = argmax\, T(\tilde{x}; \theta_T^*), \\ \tilde{x} = G(z|a; \theta_G^*)\},\end{aligned} \tag{3}$$

where T represents teacher model, θ_T^* and θ_G^* are the optimised parameters of teacher and generator, $\tilde{x}^*$ is the high-quality generated features, y^* is the corresponding class labels. **5): Training the student model:**

$$\min_{\theta_S} \| T^*(\tilde{x}^*; \theta_T^*) - S(\tilde{x}^*; \theta_S) \|_2^2 \tag{4}$$

where S and θ_S denotes the student model and its parameters.

In the white-box scenario, the gradient is imposed directly onto generated features and can massively improve the performance of the generator. As a trade-off, the gradient feedback is mid-risk information (may lead to teacher model leaking) whereas the softmax and regularization feedback are low-risk.

Black-Box Scenario. Black-box scenario only differs from the white-box scenario in the guidance provided by teacher model in the second step. Only low-risk regularization and softmax output can be requested from the teacher model so as to avoid the model leaking risk. Specifically, generated features $\tilde{x} = G(z|a;\theta_G)$ is uploaded to data owner to compute the softmax and divergence regularization. The data owner then creates a request so that the feedback can be downloaded. Generated data can validate whether its conditional class input can match the teacher softmax output and misclassified samples are filtered out. Generator G and student network S are then trained as an end-to-end model as follows:

$$\min_{\theta_G,\theta_S} \|T^*(\tilde{x};\theta_T^*) - S(\tilde{x};\theta_S)\|_2^2 + \alpha\mathcal{R}(\tilde{x}), \tag{5}$$

where θ_G, θ_S are parameters of generator and student model.

This paper mainly focus on investigating following research questions **(RQ)**: **1)** what are the impacts of different teacher feedback information on quality and diversity of generated data? **2)** different semantic information as generation condition and their impacts; **3)** trade-off between performance and security in white-box and black-box; **4)** can student generate new knowledge beyond the limitation of a quasi-omniscient teacher? **5)** previous work uses real seen data and generated unseen data, which causes bias towards seen classes. In PE-ZSL, both seen and unseen classes are trained using generated data, which improves consistency between seen and unseen classifier in GZSL problem.

3.3. Privacy-Enhanced Zero-Shot Classification

After the training process, generator can synthesize features of good quality and student network can predict class labels of test features. With the omniscient teacher, where seen and unseen classes are available, the generator can synthesize features of all classes. Given the test features, we can obtain the predicted class labels as follows:

$$y^* = \underset{y\in\mathcal{Y}}{argmax}\, p(y|x,\theta_S^*), \tag{6}$$

where θ_S^* denotes optimised parameters of student.

With the quasi-omniscient teacher, where only seen class data is available, the problem is more challenging. The generator is utilized to synthesize data of unseen classes. Given the noise z and unseen class semantic embedding, the generated features can be obtained as $\tilde{x} = G(z|a;\theta_G^*)$. Then it is converted into a supervised learning task. The generated features are adopted to train a classifier C and class labels of test features can be predicted through optimized classifier.

4. EXPERIMENTS

Datasets and Implementation Details. We evaluate our PE-ZSL model on three benchmark datasets: AWA1[5], AWA2 [25]) and aPY [26]. AWA1 and AWA2 consist of 30,475 and 37,322 images of 50 classes. aPY contains 15,539 images of 32 classes. As semantic representation, we use 768-dimensional word embedding generated by BERT [15]. Following [25], we adopt the 2048-dimensional ResNet101 features as image representation. As for data split, we follow the proposed data split in [25] for quasi-omniscient teacher. Omniscient teacher is trained with all classes, so we split unseen classes randomly into training and test sets following [21]. Student and teacher models have the same architecture, which has two hidden layers with 1024 and 512 units. Generator contains a single hidden layer with 4096 hidden units. The dimension of noise vector z is set to 20 for all datasets. The regularization term weight is set to 0.5 for AWA1 and AWA2, and 1 for aPY. The number of generated features is 400 in average per class for all datasets.

Evaluation Protocol. Following [25], we adopt the per-class average top-1 accuracy (T1) for CZSL task. We use harmonic mean $H = (2\times u\times s)/(u+s)$ for evaluation in GZSL, where u and s denote average per-class top-1 accuracy on unseen and seen classes, respectively.

4.1. Main Results

Comparisons with State-of-Arts. We present experimental results in both CZSL and GZSL task in Table 1. Considering this is the first PE-ZSL work, we provide comparison with traditional state-of-the-arts as a reference. To investigate **RQ1**, we show results under two kinds of feedback from omniscient and qusi-omniscient teacher. PE-ZSL model with omniscient teacher achieves promising performance in both CZSL and GZSL in white-box scenario. 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 black-box scenario, the accuracy on unseen classes is 4.9% higher than seen classes on AWA1. It indicates that PE-ZSL model is promising to mitigate class-level overfitting issue in GZSL task proposed in **RQ5**. Compared with inductive ZSL methods, results show that our model with quasi-omniscient teacher in white-box scenario gains satisfactory performance in GZSL especially on aPY, with 7.3% higher performance on harmonic mean. It is very impressive that student model can generate new knowledge beyond the source data of teacher model as discussed in **RQ4**. For black-box scenario, results show our PE-ZSL model outperforms random guessing, which are around 10% on AWA1, AWA2 and 8% on aPY. The white-box achieves better performance than black-box, indicating that gradient guidance provides more information.

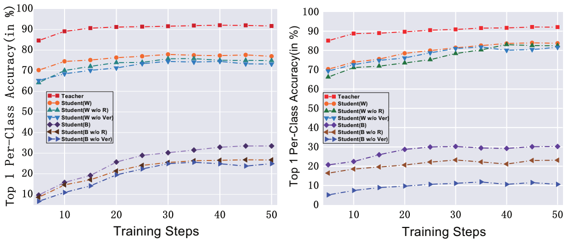
Comparisons in Black-Box Scenario. As it is the first time to propose this setting, we provide several baselines for comparison in Table 2. We provide label and attribute for conditional feature generation to investigate **RQ2**. Our proposed framework with BERT embedding achieves the best performance, *i.e.*, with 18.0% and 23.4% increases in unseen accuracies on AWA1. Results show that our framework gains

Table 1: Comparison results in CZSL and GZSL tasks. 'WB' & 'BB' represents white- & black-box scenario, '*' represents TZSL method. 'PE-ZSL+WB/BB*' and 'PE-ZSL+WB/BB' represent our model with omniscient and quasi-omniscient teacher.

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
DAP [5]	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 [10]	59.9	62.5	39.7	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEM [16]	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 [17]	68.2	-	-	57.9	61.4	59.6	-	-	-	-	-	-
CE-GZSL [18]	71.0	70.4	-	65.3	73.4	69.1	63.1	78.6	70.0	-	-	-
SDGZSL [19]	-	74.3	47.0	-	-	-	69.6	78.2	73.7	39.1	60.7	47.5
ICCE [20]	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* [21]	69.0	-	41.5	54.8	88.5	67.7	-	-	-	37.4	**87.9**	52.5
GMSADE* [22]	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* [23]	**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* [24]	-	**82.4**	49.8	-	-	-	56.2	82.2	66.7	39.3	72.9	51.0
PE-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
PE-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
PE-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
PE-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**

Table 2: Experimental results in black-box scenario with omniscient teacher in both CZSL and GZSL task.

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
PE-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**



(a) AWA1 (b) aPY

Fig. 3: Epoch analysis for unseen accuracy. 'Ver': label verification. 'R': regularization term.

obvious improvement in accuracy with label verification, *i.e.*, with 20.2% higher performance on harmonic mean on aPY dataset. And results indicate the effectiveness to adopt 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 PE-ZSL model in black-box scenario with omniscient teacher.

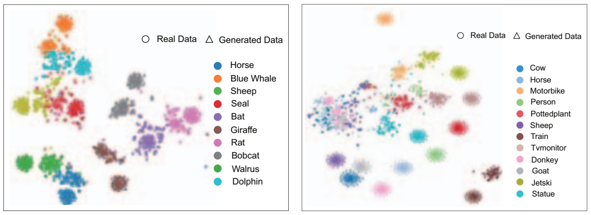
Performance vs Framework Privacy. Compared to traditional ZSL methods, the performance under white-box scenario is very promising, since data privacy is already preserved and our model can still achieve adequate performance. Compare with white-box scenario, black-box is more secure but scarifies classification performance. Thus, the performance of black-box scenario is reasonable because both data privacy and model safety are guaranteed as proposed in **RQ3**.

4.2. Analysis and Discussion

Student Performance Analysis. We present performance of teacher and student model with increasing training steps in both scenarios on AWA1 and aPY in Figure 3. Student in white-box scenario obtains results close to teacher, indicating the effectiveness of gradient guidance. Besides, results show that model achieves better performance with regularization term, indicating the effectiveness of feature distribution during training. And statistics also show that framework performs better with label verification in both scenarios, which indicates its necessity because it can mitigate the negative influence caused by generated features of inferior quality.

Table 3: Results in white-box scenario with 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



(a) AWA1 (b) aPY

Fig. 4: t-SNE visualization on AWA1 and aPY

Teacher Model Privacy Evaluation. Table 3 shows the performance with different privacy budgets ϵ when DP is added for teacher training. Detailed illustration is provided in supplementary material. $\epsilon = \infty$ indicates the non-private performance as baseline, *i.e.*, without DP for teacher training. Results show that the larger ϵ the higher performances of both teacher model and PE-ZSL model, indicating that smaller ϵ leads to more data security protection. There exists a trade-off between performance and privacy level and we can adjust the privacy budget to achieve a balance.

Quality of Generated Features. Figure 4 shows the t-SNE visualization of real and generated unseen features in both scenarios with omniscient teacher on AWA1 and aPY. We randomly choose a part of features for clear visualization. Generated features have distribution close to real ones and they are more class-level clustered, indicating effectiveness of feature generation under the supervision of teacher guidance, even though real data is unavailable. Therefore, generated features can be viewed as a suitable replacement for real features.

5. CONCLUSION

This paper has presented a privacy-enhanced ZSL paradigm via data-free knowledge transfer. A pre-trained teacher model was deployed on the data owner as data safeguard to provide guidance for model training. We extensively studied the 'black-box' and 'white-box' scenarios and their trade-off in performance and framework privacy. Our model maintains promising performance in CZSL and GZSL tasks despite the absence of real data during training. Future development of PE-ZSL can focus on model design, reducing communication cost and improving performance in the inductive setting.

6. REFERENCES

[1] Paul Voigt and Axel Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[2] Cynthia Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[3] Jakub Konevcnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[4] Guodong Xu, Ziwei Liu, and Chen Change Loy, "Computation-efficient knowledge distillation via uncertainty-aware mixup," *arXiv preprint arXiv:2012.09413*, 2020.

[5] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE TPAMI*, vol. 36, no. 3, pp. 453–465, 2013.

[6] Haoran Duan, Yang Long, Shidong Wang, Haofeng Zhang, Chris G Willcocks, and Ling Shao, "Dynamic unary convolution in transformers," *IEEE TPAMI*, 2023.

[7] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *IEEE TPAMI*, 2021.

[8] Yang Long, Li Liu, Fumin Shen, Ling Shao, and Xuelong Li, "Zero-shot learning using synthesised unseen visual data with diffusion regularisation," *IEEE TPAMI*, vol. 40, no. 10, pp. 2498–2512, 2017.

[9] Junyan Wang, Zhenhong Sun, Yichen Qian, Dong Gong, Xiuyu Sun, Ming Lin, Maurice Pagnucco, and Yang Song, "Maximizing spatio-temporal entropy of deep 3d cnns for efficient video recognition," *arXiv preprint arXiv:2303.02693*, 2023.

[10] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid, "Label-embedding for attribute-based classification," in *CVPR*, 2013.

[11] Rui Gao, Xingsong Hou, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Zhao Zhang, and Ling Shao, "Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning," *IEEE Transactions on Image Processing*, 2020.

[12] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong, "Transductive multi-view zero-shot learning," *IEEE TPAMI*, vol. 37, no. 11, pp. 2332–2345, 2015.

[13] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean, "Zero-shot learning by convex combination of semantic embeddings," in *ICLR*, 2014.

[14] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *ECCV*, 2016.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a deep embedding model for zero-shot learning," in *CVPR*, 2017.

[17] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata, "Feature generating networks for zero-shot learning," in *CVPR*, 2018.

[18] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang, "Contrastive embedding for generalized zero-shot learning," in *CVPR*, 2021.

[19] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang, "Semantics disentangling for generalized zero-shot learning," in *ICCV*, 2021.

[20] Xia Kong, Zuodong Gao, Xiaofan Li, Ming Hong, Jun Liu, Chengjie Wang, Yuan Xie, and Yanyun Qu, "En-compactness: Self-distillation embedding & contrastive generation for generalized zero-shot learning," in *CVPR*, 2022.

[21] Haofeng Zhang, Li Liu, Yang Long, Zheng Zhang, and Ling Shao, "Deep transductive network for generalized zero shot learning," *Pattern Recognition*, vol. 105, pp. 107370, 2020.

[22] Omkar Gune, Mainak Pal, Preeti Mukherjee, Biplab Banerjee, and Subhasis Chaudhuri, "Generative model-driven structure aligning discriminative embeddings for transductive zero-shot learning," *arXiv preprint arXiv:2005.04492*, 2020.

[23] Lei Zhang, Peng Wang, Lingqiao Liu, Chunhua Shen, Wei Wei, Yanning Zhang, and Anton Van Den Hengel, "Towards effective deep embedding for zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[24] Xinpeng Li, Dan Zhang, Mao Ye, Xue Li, Qiang Dou, and Qiao Lv, "Bidirectional generative transductive zero-shot learning," *Neural computing and applications*, pp. 5313–5326, 2021.

[25] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata, "Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly," in *CVPR*, 2017.

[26] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth, "Describing objects by their attributes," in *CVPR*, 2009.