Contrastive Adaptation on Domain Augmentation for Generalized Zero-Shot Side-Scan Sonar Image Classification

Yunpeng Jia, and Xiufen Ye, *Senior Member, IEEE*, Peng Li and Shuxiang Guo, *Fellow, IEEE*

***Abstract*—Predicting never-seen-before targets in underwater side-scan sonar (SSS) recognition is challenging due to limited data availability and complex environmental factors. Traditional supervised methods achieve high accuracies in standard tasks but fail to generalize in zero-shot scenarios. Recent style-based transfer learning methods for zero-shot learning (ZSL) in SSS have shown promise but suffer from unrealistic environmental and sample assumptions. To overcome these limitations, we propose Contrastive Adaptation of Domain Augmentation (CADA), a novel learning paradigm that utilizes background fusion and noise modeling to expand generalized zero-shot learning (GZSL), offering greater practicality in engineering. By integrating simulated SSS noise with fused backgrounds, our approach augments unseen classes, improves class separability, and mitigates overfitting. The contrastive adaptation further narrows domain distribution gaps while preserving critical intra-class semantic content information. Moreover, we introduce the first SSS image dataset tailored for the GZSL application. Experimental results show that CADA reaches up to 73.32% on the harmonic mean index, achieving over 20% higher accuracy than existing state-of-the-art style-based methods, highlighting its effectiveness for SSS target classification in GZSL settings. The code is https://github.com/JiaYP0433/CADA-Generalized-Zero-Shot-Side-Scan-Sonar-Image-Classification.**

***Index Terms*—Sonar Image classification, Generalized Zero-Shot Learning, Contrastive Adaptation, Domain Augmentation**

I. INTRODUCTION

DEEP convolutional neural networks (DCNN) has been widely applied in imaging-based target classification, including optic, acoustic, radar images, and so on. Traditional high-quality DCNN classification relies heavily on large datasets with with manually annotated labels. They become ineffective in practical scenarios when priori data are sparse or nonexistent. For instance, it is difficult for side-scan sonar (SSS) equipment [1] to acquire samples including expected targets due to the high costs of offshore operations and related policy restrictions. SSS imaging systems offer several advantages, including high resolution, wide coverage, depth penetration, and versatility, making them invaluable for deep-water exploration. Compared to forward-looking sonar, SSS [2] covers a wider swath of the seafloor in a single pass alongside the survey path and provides ultra-high imaging of seafloor and objects, making it particularly effective for mapping and exploration.

Zero-shot learning (ZSL) [3], [4] addresses the task of classifying classes that have never appeared in the training samples. In practical application, generalized ZSL (GZSL) [5] is more commonly used as it allows for the classification of both seen and unseen classes in the training data. At present GZSL is used in many engineering fields, such as electric power [6], diagnosis [7], flaw inspection [8] and so on. However, research on GZSL for SSS image classification is limited, even though existing SSS recognition methods [9], [10] have satisfactory accuracy.

Li et al [11] simulated the human way and developed a ZSL image classification method for SSS by synthetizing pseudo SSS images with added noise. In this experiment, remote sensing (RS) image data were taken as seen samples to train encoder-decoder models. After that, an SSS image as a style sample was encoded together with an optical image and then transformed into a pseudo SSS image by a whitening and colouring transform (WCT) reconstruction and a decoder with added noise. Other style transfer frameworks [12], [13] have been developed to generate novel SSS images for SSS classification in the setting [11]. Huang et al [14] took shipwrecks as novel targets and generated representative shipwreck pseudo samples using SSS imaging mechanism and environment. These studies have demonstrated that encoder-decoder networks based on transfer learning are reliable mediums to synthetize unseen samples for classification.

Instead of designing a new framework, we standardize the SSS learning benchmark and paradigm tailored for GZSL classification. **Fig. 1** illustrates what an autonomous underwater vehicle (AUV) hanging SSS equipment should do for the GZSL task. The motivation is based on three terms below.

This paper was supported by the National Natural Science Foundation of China (Grant No. 42276187).

Yunpeng Jia is with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin and 150006, China (e-mail: jiayunpeng@hrbeu.edu.cn).

Xiufen Ye is corresponding author and with the College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin and 150006, China (e-mail: yexiufen@hrbeu.edu.cn).

Peng Li is with the Management School, Harbin University of Commerce, Harbin and 150018, China (lipeng@hrbcu.edu.cn)

Shuxiang Guo is with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518055, China, and with the Key Laboratory of Convergence Medical Engineering System and Healthcare Technology, Ministry of Industry and Information Technology, Beijing Institute of Technology, Beijing 100081, China (e-mail: guo.shuxiang@sustech.edu.cn).

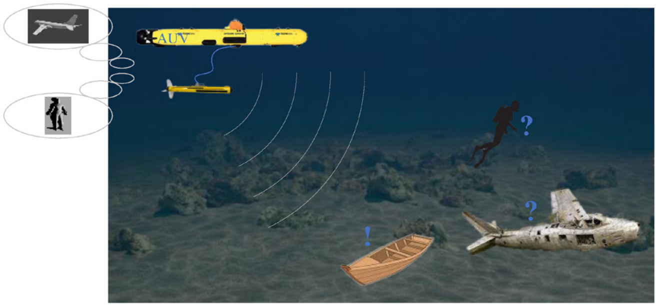
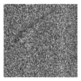


Fig. 1. Discovery: What should do the AUV hanging SSS equipment when meeting up with targets of unseen classes? Existing SSS recognition methods can easily allow the AUV host computer to identify seen targets (seabed or shipwrecks) which are common in or on water by naïve supervise learning. However, it will become difficulties when AUV encounters novel unseen targets (aircraft wreckages or human body) which are not common in the water. The purpose of GZSL is to allow AUV to recall the appearance of novel targets, and associate the general contents of the novel targets on the SSS image according to the scanned image texture.

TABLE I
REAL-WORLD SSS DATA STATISTICS FROM [15]

	Seen classes		Unseen classes	
Examples				
Class Name	Seafloor	Shipwreck	Plane-Wreckage	Person
Number	>500	>400	<100	<20

Motivation 1. There are currently no available datasets meeting with unified evaluation protocols and data splits, as suggested by the reference [5]. The split between seen and unseen classes should align with natural and cognitive principles. It means that targets of seen classes (such as seabed and shipwreck) are relatively common in underwater environment, whereas extremely rare targets are represented as unseen classes. Besides, it is crucial to ensure that unseen classes do not appear in the dataset used to pre-train the backbone network for complying with cognitive changes. This argument is supported by real-world SSS data statistics presented in Table I [15].

Motivation 2. The size of the training data is not uniformly and reasonably set. According to the GZSL criterion, the available information about unseen classes is limited to one or a few semantic features. This criterion aligns with the common understanding that the target patterns which humans recall from memory are generally rough.

Motivation 3. Although style-based transfer learning is currently the most effective way, the setting is not strictly regulated. Based on accessibility principles, style samples should not include unseen classes. Moreover, the effectiveness of style transfer is the key to ensuring the quality of GZSL.

Based on above motivations, we propose contrastive adaptation with domain augmentation (CADA) – a carefully designed paradigm that uses style transfer as a guiding principle to generate diverse pseudo-unseen samples. The CADA learning paradigm is illustrated in **Fig. 2**. Compared with directly used style transfer models [16], [17], [18], [19], CADA augments unseen pseudo SSS samples with a wide variety of bottom topographies using semantic content and background fusion, preventing classifiers from overfitting towards seen classes. Moreover, CADA narrows down the intra-class Kullback-Leibler (KL) divergence for contrastive adaptation between the content image domain and SSS image domain.

After verification across SSS image datasets, our paradigm can significantly improve classification performance using existing style transfer models for sample generation in the GZSL scenario. To summarize, our contributions are four-fold:

1) We propose a novel CADA learning paradigm for SSS image classification. To the best of our knowledge, this is the first time GZSL has been applied to underwater SSS category classification;
2) We introduce background fusion with added SSS image noise to constitute unseen class augmentation using a pre-existing style transfer framework without re-training, ensuring sample diversity and fidelity while avoiding overfitting towards seen classes;
3) We present a weighted loss function that incorporates logit adjustment and KL divergence for contrastive domain adaptation, enhancing the classification robustness and preventing model collapse;
4) We provide an available SSS dataset for GZSL research. The advancement and effectiveness of the proposed paradigm are verified, serving as an evaluation benchmark for the application of style transfer networks in the SSS image classification.

II. RELATED WORKS

Our work is closely related to the following research interests.

A. SSS Image Classification

SSS systems, mounted on a towfish or AUV, are among the most prominent sensors for underwater searches, such as unexploded ordnances, wreckages or landforms. These systems emit an acoustic ping and receive the backscattered signal while working. The recorded time-signals are processed, meanwhile, an image is formed by stacking consecutive pings on top of each other. With the breakthrough of deep learning technology, methods based on DCNN have gradually replaced traditional automatic target recognition methods [20], [21], [22].

Berthold et al [23] presented an early study on automatic sediment type classification using DCNNs, achieving 83% accuracy. Luo et al [24] combined LeNet-5 with AlexNet to handle small sample classification. Although these methods achieved high accuracy, often exceeding 90%, they are notoriously data-hungry. Compared to optic or RS imaging, SSS imaging is time consuming and technically cumbersome, making it nearly impossible to obtain a large amount of manually annotated data in advance.

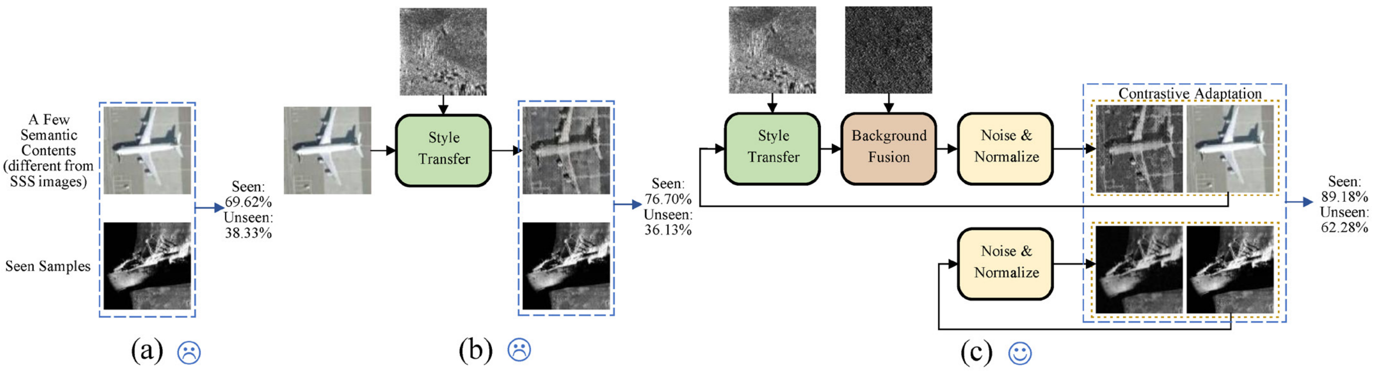

A Few Semantic Contents (different from SSS images)
Seen Samples
Seen: 69.62% Unseen: 38.33%
Style Transfer
Seen: 76.70% Unseen: 36.13%
Style Transfer
Background Fusion
Noise & Normalize
Contrastive Adaptation
Noise & Normalize
Seen: 89.18% Unseen: 62.28%
(a) ☹
(b) ☹
(c) ☺

Fig. 2 (a) paradigm directly feeding DCNN semantic contents and seen samples; (b) paradigm using style transfer models; (c) our proposed paradigm. Although the effectiveness of the paradigm (b) in the ZSL setting has been verified, it leads to overfitting when both seen and unseen classes need to be classified. The proposed paradigm achieves a satisfactory result in the GZSL scenario.

To address this problem, few-shot [25], [26] or zero-shot [11], [27] approaches are developed using transfer learning. Qin et al [28] extended these previous works by pre-training on grayscale CIFAR-10 images and augmenting data using generative adversarial networks (GANs). Transfer-learning-based recognition [1], fine-tuning backbone networks [29] and enlarging dataset by a GAN [10], [30] are among the most effective techniques for classification of a bounded number of SSS samples.

Zhao et al [31] simulated the imaging mechanism and image characteristics of SSS in marine environment to generate shipwreck samples by a target-to-target and background-to-background style transfer model. The prompt fine-tuning [32] based on vision-language models has been used for the zero-sample problem of SSS in the open-set recognition. Jiao et al [33] proposed a framework for open-set recognition tasks with long-tail SSS distribution.

In this study, we attempt to build a novel paradigm directed against SSS image classification in the GZSL scenario using style transfer.

B. GZSL

The goal of ZSL is how to make a machine imitate humans to recognize unseen classes aided by auxiliary information. Early works focused on predicting outputs from intermediate layer between feature and label spaces, usually using attributes [4] or word vectors [34] as auxiliary semantic information. GZSL [5] extends ZSL by recognizing both seen and unseen classes.

In terms of algorithms, GZSL can be broadly categorized into two major approaches: mapping-based [35] and data-based generative [36] methods. Mapping-based methods aim to map semantic representations and predicted visual representations into a shared space where the distribution of the two types of representations are similar, and intra-class distribution divergence is minimized. Improvements in mapping-based techniques, such as out-of-distribution detection [37], meta-learning [38], long-tail rebalancing [39], and attribution-transformer-based augmentation [40] have enhanced robustness against limited data and impure semantic information.

Data-based generative methods, on the other hand, offer better generalization and handle projection bias more effectively than mapping-based methods. Techniques involving pseudo features generated by GANs with variational autoencoders [41], semantic irrelevant disentanglement [42] and contrastive embedding [43] almost occupies a dominant position. To mitigate the impact of low-quality feature or semantic features, contrastive representation optimization [44] and prototype-guided sub-representation generation [45] are leading in GZSL research, both theoretically and practically.

In our approach, we use an existing style transfer network as a data-based generation model, incorporating a proposed loss function with logit adjustment to achieve minimal mapping between source and target spaces. As for semantic information of unseen SSS classes, we use a few RS or shadow photography views with content relatively close to SSS images, as they capture targets from an overhead perspective. Although the prior information of unseen classes primarily contains the target outlines, many of the leading GZSL methods effectively utilize heterogeneous transfer learning [46] to solve recognition tasks in the target domain using information from another domain.

C. Style Transfer

Johnson et al [47] first pioneered real-time transfer style from random images. WCT (Whitening and Coloring Transform) [48] whitened and then re-colored auto-encoded feature maps by singular value decomposition. PhotoWCT [49] improves on WCT by replacing the up-sampling layers with un-pooling layers that preserve spatial information, addressing distortions of object boundaries caused by WCT. Afterwards LiWCT [11] modified PhotoWCT by filling the zeros of traditional un-pooling layers with random noise values to reduce the checkerboard effect and generate various styles closer to real SSS images. Stytr2 [18] based on attention-based transformers handles long-term dependencies and retain target details. CCPL [19] simplified the feature transformation module based on [16] and [17], devising contrast coherence preservation based on the assumption of similarity between adjacent areas of the image.

The textures of SSS images are very different from those of other style images shown in **Fig. 3**. Therefore, our tailor-made GZSL paradigm for SSS classification falls within the scope

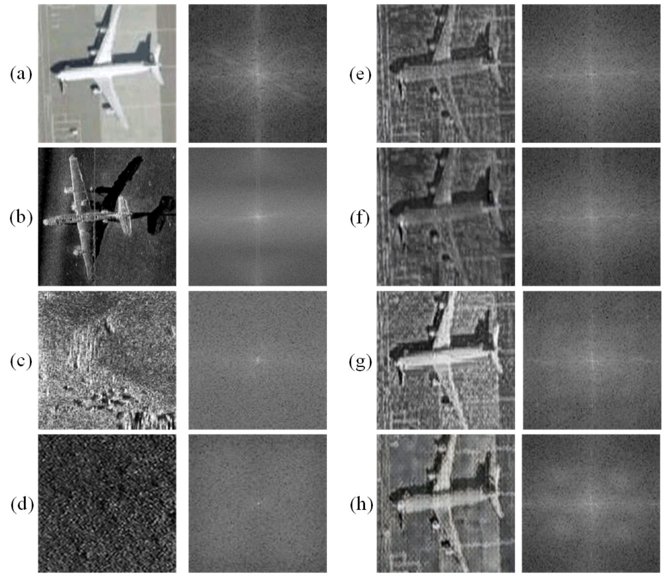


Fig. 3. Comparative analysis on images on the left and corresponding Fourier transform images on the right. (a) an RS airplane image; (b) an SSS airplane image; (c)-(d) SSS seafloor images; (e)-(h) an image generated by PhotoWCT, LiWCT, StyTr and CCPL transfer models using (c) as the style image, respectively; The SSS image texture has heterogeneous components compared to RS image texture because the central rays of SSS Fourier images are weaker but the high-frequency components in the non-central area are larger than RS. And the spatial distribution of SSS peripheral high-frequency components is more uniform than that of RS.

of heterogeneous transfer [50]. In this paper, we investigate the GZSL performance of the proposed paradigm based on existing style transfer models: PhotoWCT, LiWCT, StyTr, and CCPL. We provide reasonable explanations for the classification effects using these models, along with guidance on which transfer model is most suitable for SSS classification. The proposed paradigm, using background fusion and noise addition, takes the advantage of distilling target-invariant content akin to mapping-based methods and generates various pseudo samples for contrastive augmentation of unseen classes like data-based generative methods.

III. Preliminaries

A. Mathematical Formulation

We will provide fundamental notations for GZSL in SSS classification. Given K seen classes and L unseen classes, the sets of seen classes $\mathcal{Y}^{\mathrm{s}}$ and unseen classes $\mathcal{Y}^{\mathrm{t}}$ are disjoint, i.e. $\mathcal{Y}^{\mathrm{s}}\cap\mathcal{Y}^{\mathrm{t}}=\emptyset$. We denote $\mathcal{Y}^{\mathrm{s}}=\{1, \cdots, K\}$ and $\mathcal{Y}^{\mathrm{t}}=\{K+1, \cdots, K+T\}$.

The available training information includes seen prior data $\boldsymbol{\mathcal{D}}^{\mathrm{s}}=\{\mathcal{D}_1, \cdots, \mathcal{D}_K\}$ and target prior data $\boldsymbol{\mathcal{D}}^{\mathrm{t}}=\{\mathcal{D}_{K+1}, \cdots, \mathcal{D}_{K+T}\}$, where $\mathcal{D}_k=\{(I_{s,i}, k)|I_{s,i}\in\mathcal{I}, k\in\mathcal{Y}^{\mathrm{s}}\}_{i=1}^{N_k}$ and $\mathcal{D}_t=\{(I_{c,i}, t)|I_{c,i}\in\mathcal{C}, t\in\mathcal{Y}^{\mathrm{t}}\}_{i=1}^{N_t}$. Here, $I_{s,i}$ is an SSS image sample from seen classes, $I_{c,i}$ is a semantic content image different from SSS domain, $\mathcal{I}$ represents the SSS image space and $\mathcal{C}$ represents the other content domain space instead of the SSS domain. We assume $N_t \ll N_k, t\in\mathcal{Y}^{\mathrm{t}}\ and\ k\in\mathcal{Y}^{\mathrm{s}}$ is to comply with the GZSL criterion. Very few source samples $I_{c,i}\in\mathcal{C}$ correspond to semantic information of unseen classes. Attributes, as in [4], are generated by an infinite relational model [51] that explores similar relations between kinds of entities in each set.

Our method eliminates the need for hand-collected attribute descriptions and supervised clustering. Our goal is to learn general visual classifiers across all classes $f_{\mathrm{gzsl}}: \mathcal{I}\longrightarrow\mathcal{Y}^{\mathrm{s}}\cup\mathcal{Y}^{\mathrm{t}}$.

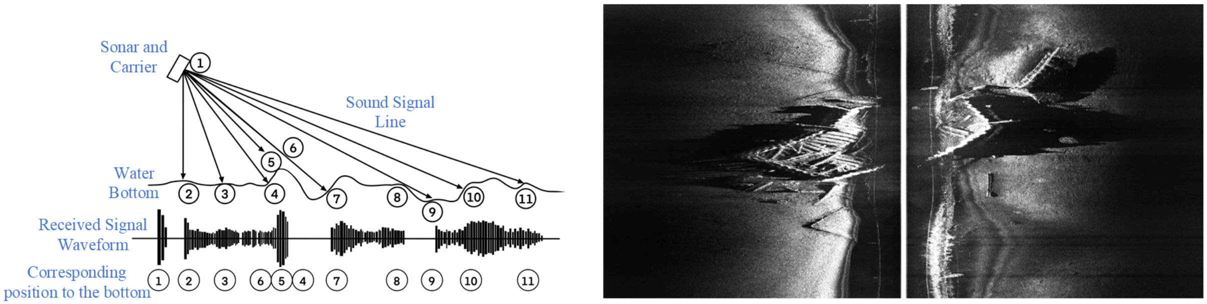
B. SSS Image Characteristic

Each side of an SSS system has a transducer array at work. It emits a short sound pulse that scatters upon hitting the seabed or objects in the water. Some of the scattered sound returns to the transducer along the original propagation route and is converted into an electrical pulse signal or ping.

The signal intensity diagram of one side of the transducer array is shown in **Fig. 4 (a)**. Region ① is located near the transducer signal receiving and transmitting region, presenting a strong received signal. The water column between regions ① and ② barely receives any signal. Region ② is directly below the transducer and also exhibits a strong signal. Additionally, regions with smooth surfaces, raised surfaces close to the transducer, and areas perpendicular to the direction of signal propagation, such as ④, ⑤, ⑥ and ⑩, also show strong signals. Conversely, the regions between ⑥ and ⑦, and between ⑧ and ⑨ are obstructed and present minimal signals. A ping signal is numerically converted by a carried processor according to its strength, producing a line of images. As the carrier moves forward, this process of the transducer transmitting and receiving sound signals is repeated, forming a converted ping signal. These signal pings are then spliced together to obtain a complete SSS image, as illustrated in **Fig. 4 (b)**.

According to grayscale distributions of images, the region with strong signal is referred to the SSS image highlighted area, the region with minimal signal is called the SSS image shadow area, and the remaining area is the SSS image background area. The highlighted area usually includes objects suspended or shallowly buried on the seabed, or be raised seabed hills. The shadow area indicates the occluded part containing only scattering noise, while the background represents broad seabed. In this paper, we disregard the influence of slant range and navigation speed on the image grayscale, assuming that all given images are preprocessed by relevant algorithms [52], [53], [54]. This assumption ensures that the background grayscale distribution is approximately uniform in pixel space and that the peak value of the grayscale histogram in the image background is approximately equal to the median value of the image.

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2025.3551028



(a) (b)

Fig. 4. (a) The returned signal intensity diagram of one side of SSS array. (b) Example of SSS image.

C. Image Style transfer

The goal of image style transfer is to render a content image using a reference style image. Style transfer models typically consist of an encoder that processes the content image, a decoder that generates the transferred image, and a transform module that handles the deep features of both the content and style images. We review four typical style transfer models, PhotoWCT, LiWCT, StyTr, and CCPL.

In PhotoWCT, the transformation module first whitens a content feature f_{o} by:

$$\hat{f}_{\mathrm{o}}=E_{\mathrm{o}}D_{\mathrm{o}}^{-\frac{1}{2}}E_{\mathrm{o}}^{\mathrm{T}}f_{\mathrm{o}}, \tag{1}$$

where D_{o} is a diagonal matrix with the eigenvalues of $f_{\mathrm{o}}f_{\mathrm{o}}^{\mathrm{T}}$ on its principal diagonal. The transferred feature $\hat{f}_{\mathrm{os}}$, which retains both whitened content and style information, is then computed by

$$\hat{f}_{\mathrm{os}}=E_{\mathrm{s}}D_{\mathrm{s}}^{-\frac{1}{2}}E_{\mathrm{s}}^{\mathrm{T}}\hat{f}_{\mathrm{o}}+m_{\mathrm{s}}, \tag{2}$$

where D_{s} is a diagonal matrix with the eigenvalues of $f_{\mathrm{s}}f_{\mathrm{s}}^{\mathrm{T}}$ (style feature f_{s}) on its diagonal, and m_s is the mean vector of f_{s}. Unlike WCT, PhotoWCT incorporates un-pooling layers to preserve spatial information in each encoder-decoder block. During training, PhotoWCT uses pixel reconstruction loss and feature reconstruction loss [47].

Building on PhotoWCT, LiWCT introduces random noise into the un-pooling layers to mitigate the checkerboard effect, generating pseudo-images with textures that more closely resemble real SSS images.

StyTr adopts transformer-based frameworks for encoding content and style images. Positional encoding $\mathcal{P}_{\mathcal{CA}}$ is learnable to acquire structural information. In the transform module, the content sequence $Z_{\mathrm{c}}=\mathcal{E}_{\mathrm{c}}+\mathcal{P}_{\mathcal{CA}}$ is first encoded into queries $Z_{\mathrm{c}}W_{\mathrm{q}}$, keys $Z_{\mathrm{c}}W_{\mathrm{k}}$, and values $Z_{\mathrm{c}}W_{\mathrm{v}}$ by a transformer encoder, where W_{q}, W_{k} and W_{v} are encoder weights. Multi-head attention is further calculated to get the encoded content sequence Y_c. The style sequence Y_s is used to generate queries $(Y_{\mathrm{c}}+\mathcal{P}_{\mathcal{CA}})W'_{\mathrm{q}}$, keys $Y_{\mathrm{s}}W'_{\mathrm{k}}$, and values $Y_{\mathrm{s}}W'_{\mathrm{v}}$ by a transformer decoder, where W'_{q}, W'_{k} and W'_{v} are decoder weights. The output sequence is further refined by multi-head attention in the transformer decoder. Compared with PhotoWCT and LiWCT, StyTr significantly improves the fusion effect of style image textures, especially in the background area, as shown in **Fig. 3**. Its transformer-based architecture and style perceptual loss enhance the preservation of content details.

CCPL features a transformation module comprising a content network $cnet(\cdot)$, a style network $snet(\cdot)$ and a convolution $conv(\cdot)$. Initially, a content feature f_{o} and a style feature f_{s} are centered along the spatial dimension, and further standardized to obtain $\hat{f}_{\mathrm{o}}$ and $\hat{f}_{\mathrm{s}}$, respectively. Then $cnet(\hat{f}_{\mathrm{o}})\to\hat{f}_{\mathrm{o}}$ and $snet(\hat{f}_{\mathrm{s}})\to\hat{f}_{\mathrm{s}}$ are flattened along the spatial dimensions. The transferred feature $\hat{f}_{\mathrm{os}}$ is computed by

$$\hat{f}_{\mathrm{os}}=\hat{f}_{\mathrm{s}}\otimes\hat{f}_{\mathrm{s}}^{\mathrm{T}}\otimes\hat{f}_{\mathrm{o}}, \tag{3}$$

where $\otimes$ denotes matrix multiplication along the channel dimension. The transformed feature $\hat{f}_{\mathrm{os}}$ is further refined by $conv(\hat{f}_{\mathrm{os}})+m_{\mathrm{s}}$, where m_s is the spatial mean of $\hat{f}_{\mathrm{s}}$. CCPL also includes the style perceptual loss.

CCPL has two key characteristics. **1.** It preserves content coherence within the neighborhood of pixels because the distance between the vector of any pixel and the one of its neighboring pixels is smaller than the distance between vectors at the other pixels in the pixel domain. **2.** According to Eq. (3), the synthesized images contain more texture details about the target domain in the regions of interest with higher values of $\hat{f}_{\mathrm{o}}$.

IV. Methodology

In this section, we elaborate on our proposed SSS classification paradigm for GZSL classification, as illustrated in **Fig. 2(c)**. The paradigm comprises three main components: background fusion, SSS noise modeling and contrastive logit adjustment loss. We give the stepwise pseudocode description of our CADA paradigm. Background fusion is applied exclusively to $I_c\in\mathcal{C}$ for unseen classes, while every SSS style sample $I_s\in\mathcal{I}$ is sourced from seen training samples.

A. Background Fusion

To diversify the generated samples, we use a background fusion unit. We assume that the grayscale distribution of generated images is relatively uniform compared to that of SSS style images. Especially after training with style perceptual loss in StyTr and CCPL, there has the relation (about spatial means $\mu(\cdot)$ and standard deviations $\sigma(\cdot)$):

$$\mu(I_{\mathrm{g}})\approx\mu(I_{\mathrm{s}}),\ \sigma(I_{\mathrm{g}})\approx\sigma(I_{\mathrm{s}}), \tag{4}$$

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2025.3551028

Algorithm 1 Pseudocode of the CADA Algorithmic Core Parts

```
1. For the processing of each sample in the dataset before training:
   a. if the sample's label is in the unseen classes:
      i. Transfer the style of the sample image by a style transfer network.
      ii. Clip a reference image to range [0, 1] and apply random scaling.
      iii. Calculate a background mask by Eq (5).
      iv. Fuse the sample with the reference image and the mask by Eq (6).
   b. Add Rayleigh and Gaussian noise to the sample image by Eq. (7).
   c. Normalize the image to ensure all the values are between 0 and 1.

2. For each training step:
   a. Feed the original sample and the processed sample into the classifier.
   b. Calculate the cross-entropy loss by Eq. (13).
   c. Compute the KL divergence loss by Eq. (17).
   d. Calculate the total loss by Eq. (18).
   e. Perform backpropagation to update the model's parameters.
```

where I_{g} and I_{s} represent the generated image and the style image features, respectively. The grayscale histogram distribution of real SSS images is inevitably biased due to sharp noise and rugged texture. For simplicity, we utilize the grayscale median as an estimator based on the peak characteristics of the gray distribution.

We compute a background mask by

$$I_{\mathrm{mask}}=\exp\left(-\left|I_{\mathrm{g}}-med_{\mathrm{g}}\right|*\lambda\right), \tag{5}$$

where med_{g} is the gray median of I_{g} and λ is a mask factor. A smaller gray value of $\left|I_{\mathrm{g}}-med_{\mathrm{g}}\right|$ indicates a greater similarity to the background area. Given a reference image I_{r} randomly extracted from seen classes that contains no objects, we obtain a fusion output I_{o} as:

$$I_{\mathrm{o}}=I_{\mathrm{g}}\cdot(1-I_{\mathrm{mask}})+I_{\mathrm{r}}^{\gamma}\cdot I_{\mathrm{mask}}, \tag{6}$$

where γ is a coefficient randomly selected from the interval [0.9, 1.1] . The gamma correction of I_{r} facilitates the diversification of the generated background distribution. To ensure that the gray values in I_{o} are within the range [0, 1], both I_{g} and I_r are normalized so that their grayscale values are confined within [0, 1].

B. Noise Modelling

During sound wave transmission, interactions with the water body and seabed result in multipath reflections, refraction, and reverberation [55], leading to noise in SSS imaging. To simulate these effects, we construct a noise modeling item that incorporates multiplicative noise, which correlates highly with signal strength, to emulate echo interference. The multiplicative noise approximately follows the Rayleigh distribution [56]. In addition, we introduce additive noise, which is independent of signal strength and follows a Gaussian distribution [57] to model imaging disturbances caused by the transducer.

Based on the noise modeling, the fusion output I_{o} for unseen classes (the original SSS image I_{s} for seen classes) is transformed into I_{on} by

$$I_{\mathrm{on}}=\begin{cases}\|(N_{\mathrm{mul}}+1)\cdot I_{\mathrm{s}}+N_{\mathrm{add}}\|, \text{for seen class samples}\\ \|(N_{\mathrm{mul}}+1)\cdot I_{\mathrm{o}}+N_{\mathrm{add}}\|, \text{for unseen class samples}\end{cases}, \tag{7}$$

where N_{mul} and N_{add} are the images of multiplicative noise and additive noise, respectively, both of which are with the same size as I_{o}. And $\|\cdot\|$ normalizes the values to the range [0, 1].

C. Logit Adjustment

In GZSL classification, the sample distribution across classes is imbalanced due to the incomplete and limited prior information of unseen classes. Existing classifiers typically optimize global accuracy by modeling the posterior probability $q(y_x|I_x)$:

$$A_{\mathrm{g}}=\mathbb{E}_{I_x\sim p(I_x)}q(y_x|I_x), \tag{8}$$

where $p(I_x)$ is a uniform distribution over all data, and y_x denotes the class label for the image I_x. However, Eq. (8) neglects the sample distribution imbalance between seen and unseen domains. To address this, we introduce an adjusted accuracy A_{adj}:

$$A_{\mathrm{adj}}=\mathbb{E}_{I_x\sim p(I_x)}\frac{q(y_x|I_x)}{p(\mathcal{Y}|y_x)}, \tag{9}$$

where $p(\mathcal{Y}|y_x)$ represents the seen-unseen prior probability. We define a network $f(\cdot)$ composed of a DCNN and a classifier in series, which of the output is a logit vector. Based on deduction theory [39], we build the adjustment for the logit as follow:

$$\frac{q(y_x|I_x)}{p(\mathcal{Y})}\propto \exp f(I_x)_{y_x}, \tag{10}$$

where $f(I_x)_{y_x}$ represents the logit value of the image I_x assigned to class y_x. After applying SoftMax normalization, we compute adjusted $q(y_x|I_x)$ by:

$$q(y_x|I_x)=\mathrm{Softmax}\left(f(I_x)+\log\left(p(\mathcal{Y}|y_x)\right)\right)_{y_x}. \tag{11}$$

We deploy the prior probability $p(\mathcal{Y}|y_x)$ based on the frequency of seen and unseen domains as:

$$p(\mathcal{Y}|y_x)=\begin{cases}\alpha/(1+\alpha)\,, y_x\in\mathcal{Y}^{\mathrm{s}}\\ 1/(1+\alpha)\,, y_x\in\mathcal{Y}^{\mathrm{t}}\end{cases}, \tag{12}$$

where α is a ratio and the setting will be analyzed in detail in the experimental section.

D. Specific Training Process

During training, we adopt a contrastive loss function that incorporates the KL divergence between training samples (I_x, y_x) and their processed counterparts I_{on}. Specifically, the training samples (I_x, y_x) come from either the seen classes (i.e. $(I_x, y_x)\in\mathcal{D}_k$ for $k\in\mathcal{Y}^{\mathrm{s}}$) or the unseen classes (i.e. $(I_x, y_x)\in\mathcal{D}_t$ for $t\in\mathcal{Y}^{\mathrm{t}}$). The contrastive loss $\mathcal{L}_{\mathrm{ce}}$ is:

$$\mathcal{L}_{\mathrm{ce}}=-\left(\mathbb{E}\left[\log\left(q(y_x|I_x)\right)\right]+\mathbb{E}\left[\log\left(q(y_x|I_{\mathrm{on}})\right)\right]\right). \tag{13}$$

Let the classifier be represented by a function $f(\cdot)$. To reduce the distribution distance between samples I_x and I_{on} in the class feature space, the predicted probability distribution vectors $prob_x$ and $prob_{on}$ are computed. The probability for the class index y is represented as:

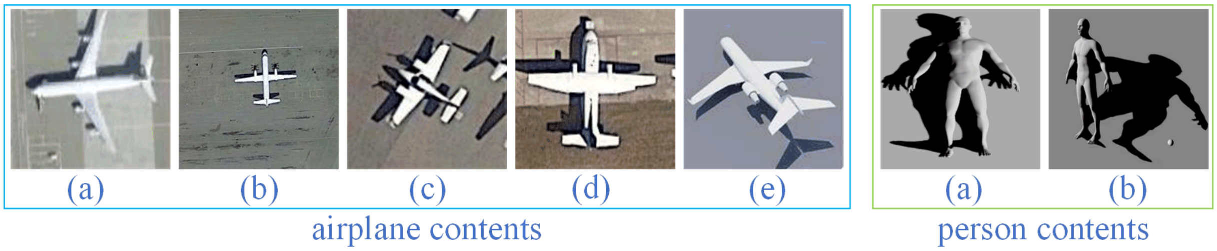


Fig. 5. Semantic content image instances for the two unseen classes, airplane and person. These indices are given.

TABLE II
GZSSS DATA STATISTICS. TR: TRAINING, TS: TESTING, AND C: CONTENT.

Domain	Seen		Unseen	
Class name	Seabed	Ship	Airplane	Person
Tr number	345	292	0	0
Ts number	233	195	60	14
C number	0	0	5	2

TABLE III
COMPARISON ACCURACY (%) WITH THESE BASELINES AND THEM ADDED WITH THE PROPOSED PARADIGM (+CADA). THE BEST RESULT IS IN BOLD

	As	*Au*	*H*
SoftMax	69.62±0.12	38.33±0.20	49.44±0.19
PhotoWCT	56.63±2.93	38.86±3.61	45.92±2.65
LiWCT	54.61±5.41	39.36±2.51	45.48±1.43
FastNoise	50.56±6.17	42.61±3.34	46.25±3.12
StyTr	75.26±2.74	35.97±2.33	48.61±2.11
GCEA	79.36±2.28	39.31±2.42	52.58±2.09
CCPL	76.70±0.76	36.13±1.16	49.11±1.10
SoftMax+CADA	77.91±2.02	44.94±1.94	56.97±1.71
PhotoWCT+CADA	74.73±1.97	49.56±1.51	59.57±1.33
LiWCT+CADA	69.60±2.31	53.23±3.01	60.26±2.01
FastNoise+CADA	73.39±4.43	51.99±4.64	60.86±4.08
StyTr+CADA	82.02±3.48	**62.78**±2.52	71.05±1.84
GCEA+CADA	86.76±1.73	54.48±2.66	66.93±2.53
CCPL+CADA	**89.18**±0.57	62.28±1.81	**73.32**±1.26

$$prob_x(y)=\text{Softmax}\left(f(I_x)+\log\left(p(\mathcal{Y}|y)\right)\right), \tag{14}$$

$$prob_{on}(y)=\text{Softmax}\left(f(I_{on})+\log\left(p(\mathcal{Y}|y)\right)\right), \tag{15}$$

The KL divergence between I_x and I_{on} is then expressed as:

$$\text{KL}(prob_x||prob_{on})=\sum_i prob_x(i)\log\left(\frac{prob_x(i)}{prob_{on}(i)}\right), \tag{16}$$

The KL divergence term in the contrastive loss function is given by:

$$\mathcal{L}_{\text{kld}}=\mathbb{E}[\text{KL}(prob_x\|prob_{on})]+\mathbb{E}[\text{KL}(prob_{on}\|prob_x)], \tag{17}$$

The total loss for the proposed CADA learning paradigm is:

$$\mathcal{L}_{total}=\mathcal{L}_{\text{ce}}+\mathcal{L}_{\text{kld}}*\frac{(1-epoch)}{epoch_{\text{max}}}, \tag{18}$$

where *epoch* is a training epoch number and $epoch_{\text{max}}$ is the maximum epoch. The core parts of the CADA pseudocode are provided in **Algorithm 1**.

The KL divergence weight decreases linearly with increasing epochs. This approach aims to minimize the distribution discrepancy between SSS images and other content domain images during the early training phase, while later focusing on learning knowledge from real samples.

V. EXPERIMENTS AND ANALYSIS

A. Datasets

There are no publicly available SSS image datasets specifically for GZSL scenario, though KLSG [2], SCTD [22] and LZSSS [11] each catering to different applications. Notably, LZSSS only contains two categories: airplane and ship (shipwreck) for ZSL. KLSG contains three categories airplane, ship and person (frogmen) for target detection. SCTD contains seabed (seafloor), airplane and ship for transfer classification.

To define our dataset for the GZSL scenario, we first determine the setting criteria for seen classes in the light of the motivation 1-2. First, the chosen classes, seabed and ship, are prevalent underwater objects, with the seabed being especially common. Second, similar items to ships, such as aircraft carriers (n02687172), catamarans (n02981792), and container ships (n03095699), are present in ImageNet, a widely used dataset for pre-training [5]. Other types of targets are considered as unseen classes.

We provide a data benchmark GZSSS, which includes two seen classes — seabed and ship, and two unseen classes — airplane and person for GZSL. The seabed images are sourced entirely from SCTD, while other samples were curated from the datasets, removing duplicates and reintegrating them to create a comprehensive set.

The statistics of GZSSS are summarized in Table II. For the seen classes, the dataset is split into 60% for training and 40% for testing. All samples of the unseen classes are used exclusively for testing. Knowledge for the unseen classes is trained using only a limited number of content images, as illustrated in Fig. 5, rather than SSS images. Specifically, four airplane content images come from optical RS imaging, while the remaining one is from photographic imaging. Both person content images are sourced from photographic imaging in [27]. These content images cover a broad range of morphological and semantic information.

B. Implementation Details

Our experiments utilized ResNet50 after pre-trained on ImageNet as the backbone, followed by a linear layer added as the classifier with an output dimension of 4 (corresponding to the total number of classes). We employed the pre-trained CCPL as the style transfer framework without additional retraining.

During the training phase, we used a Reduce-LR-On-

TABLE IV
MEAN HARMONIC MEAN ACCURACY AND GAIN (%) OF ABLATION RESULTS FOR THE PROPOSED PARADIGM COMBINED WITH THE FOUR STYLE TRANSFER MODELS

	PhotoWCT	LiWCT	StyTr	CCPL
-	45.92	45.48	48.61	49.11
w/ dual	48.98 (3.06 ↑)	51.68 (6.20 ↑)	69.75 (21.14 ↑)	64.02 (14.91 ↑)
w/ dual&bf	49.62 (3.70 ↑)	48.84 (3.36 ↑)	64.95 (16.34 ↑)	68.15 (19.04 ↑)
w/ dual&n	57.95 (12.03 ↑)	57.45 (11.97 ↑)	66.95 (18.34 ↑)	71.34 (22.23 ↑)
w/ dual&bf&n	57.78 (11.86 ↑)	58.66 (13.18 ↑)	67.62 (19.01 ↑)	71.61 (22.50 ↑)
w/ dual&kld	50.53 (4.61 ↑)	50.88 (5.40 ↑)	69.91 (21.30 ↑)	66.15 (17.04 ↑)
w/ dual&bf&n&kld	**59.57** (13.65 ↑)	**60.26** (14.78 ↑)	**71.05** (22.44 ↑)	**73.32** (24.21 ↑)

Plateau optimizer with a patience parameter of 2. The learning rates were set to 10^{-3} for the backbone and 2×10^{-3} for the classifier. The batch size was 64. For balanced domain sampling, we randomly selected one content image for each of the two unseen classes and 31 training samples for each of the two seen classes. The maximum number of epochs was set to 50. All input images were converted to grayscale, ensuring that the three-color channels were identical, and normalized with a mean of (0.3628, 0.3628, 0.3643) and a standard deviation of (0.1500, 0.1500, 0.1505). Data augmentation was performed to expand the training and testing samples to four times their original size by applying horizontal and vertical flips, as well as both simultaneously.

The logit adjustment ratio α in Eq. (12) was set as 150, 350, 200 and 100 for the style generative models PhotoWCT, LiWCT, StyTr, and CCPL, respectively. The Rayleigh distribution scale for multiplicative noise and the Gaussian distribution scale for additive noise were set to default values of 0.01 and 0.1, respectively. Detailed discussions on these hyperparameters are provided in the Parameter Analysis section.

The accuracies of average seen classes As and average unseen classes Au are calculated based on the universal evaluation protocols [5]. The simultaneous classification accuracy of both seen and unseen classes is evaluated by a harmonic mean as:

$$H = 2 * As * Au/(As+Au)\,. \tag{19}$$

This harmonic mean H is the most crucial criterion to assess the GZSL performance. We repeated the experiments 10 times with different random seeds for Sections V.C-D and H, and conducted additional experiments for each variant in Sections V.E-G for further analysis.

C. Performance Results

To showcase the effectiveness of our proposed method, we compared it against several baselines, including SoftMax, which directly uses DCNN outputs with samples, as shown in **Fig. 2(a)**; Softmax+CADA, which incorporates our proposed paradigm without style generative models as illustrated in **Fig. 2(c)**; and style transfer models used directly for training, depicted in **Fig. 2(b)**. The proposed CADA was integrated with style generative models. We recorded the arithmetic mean $\pm$ population standard deviation of As, Au and H from 10 runs using different random seeds and the results are presented in Table III.

The results reveal that our proposed learning paradigm significantly outperforms the methods that do not use the proposed paradigm and only uses the generative model in terms of Au. Specifically, the CADA generative model combined with the proposed paradigm outperforms the CCPL baseline by 24.21% in terms of H, achieving the best performance in As and H. It demonstrates that its superior effectiveness in SSS image classification with style-transfer-based strategies compared to direct training approaches.

CCPL mitigates the limitations of directly applying style transfer models compared to direct SoftMax in GZSL scenarios. And, CCPL shows greater stability than the other style transfer methods, including LiWCT and FastNoise [27], which have been effective by adding random noise to the feature space. Although the results obtained by directly using CCPL to transfer the model are not as good as the latest GCEA model [13], CCPL still outperforms it when combined with the proposed learning paradigm. Thus, the paradigm based on the improvements seen in CCPL is the most viable, with StyTr performing second best.

D. Ablation Results

We conducted a series of experiments to assess the impact of individual components within the CADA framework. The components include dual-loss ('dual'), which combines the original sample images or unseen content images with the processed images, background fusion ('bf'), noise modeling ('n') and KL divergence loss ('kld'). Table IV presents the mean H results and performance gains for the four generative models, PhotoWCT, LiWCT, StyTr and CCPL.

The results show that while LiWCT generally underperforms compared to PhotoWCT, it benefits significantly from the inclusion of dual-loss or KL divergence loss. This is attributed to the random noise addition in the un-pooling layers, which helps the DCNN recognize subtle textures and augments the features of non-interesting background areas. Consequently, LiWCT integrated with CADA outperforms PhotoWCT. In contrast, PhotoWCT faces issues with biased content representation, causing learned domain knowledge to deviate from the target domain, especially after contrastive adaptation.

When only the dual-loss component is applied, StyTr and CCPL, both of which excel in content retention, outperform PhotoWCT and LiWCT when integrated with CADA. Notably, StyTr with only the dual-loss component performs better than StyTr with additional background fusion and noise modeling. This suggests that the SSS texture is effectively embedded into the image background regions using StyTr, and additional components might lead to overcorrection. However,

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2025.3551028

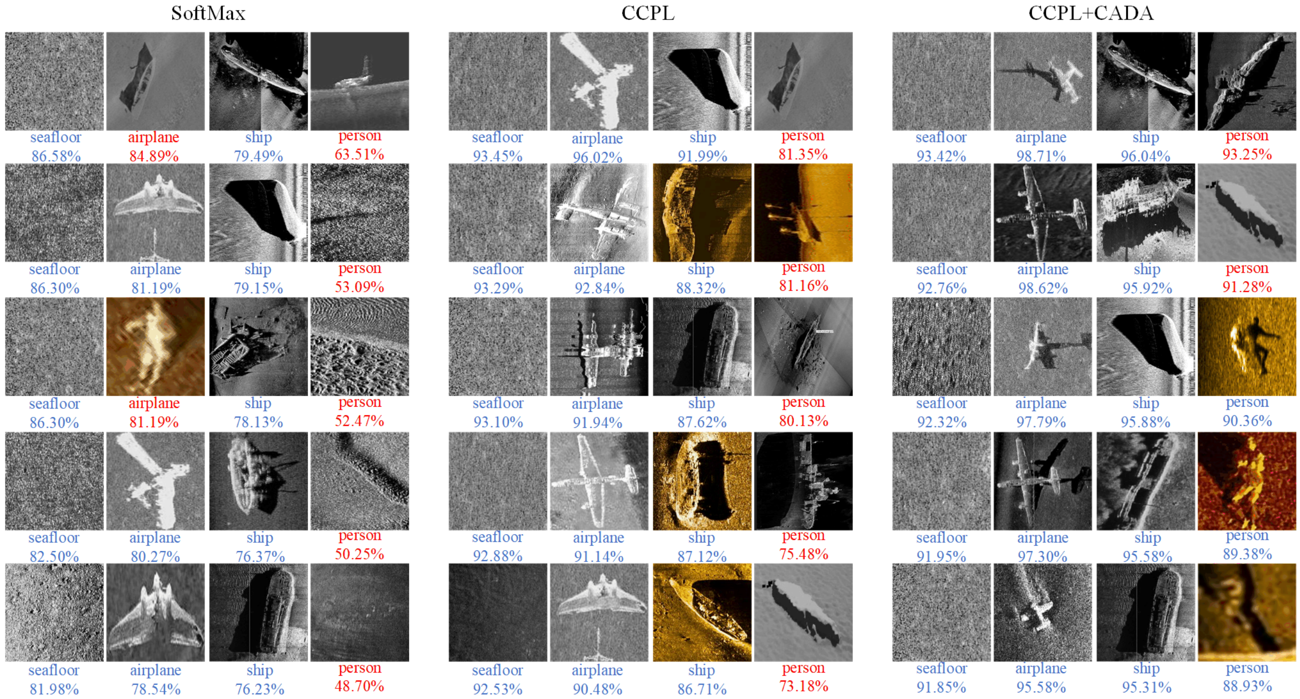


Fig. 6. The top 5 instances of each classification confidence about SoftMax, CCPL and CCPL+CADA are displayed. The prediction names and confidence scores are shown below each picture. Correct results are marked in blue and incorrect ones are in red.

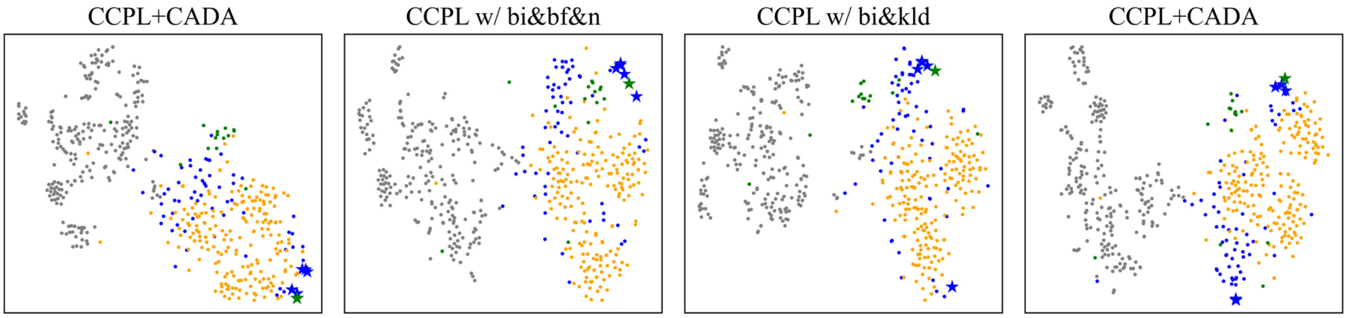


Fig. 7. t-SNE visualization results with comparison among the original CCLP, two ablation approaches and the proposed paradigm. Circle points ● and big five-pointed stars ★ represent sample features and content features, respectively. Grey (●), orange (●), blue (● ★), and green (● ★) denote the classes of seafloor, ship, airplane and person, respectively.

the best results are achieved by incorporating both background fusion and noise modeling under contrastive adaptation constraints. This indicates that contrastive adaptation augmentation mitigates domain bias effectively, regardless of the quality of the synthesized SSS texture.

For CCPL, CADA fully exploits its advantages in maintaining content coherence and achieving realistic texture blending in regions of interest. CADA successfully combines the strengths of each component, making CCPL the optimal generative model.

E. Instance Visualization

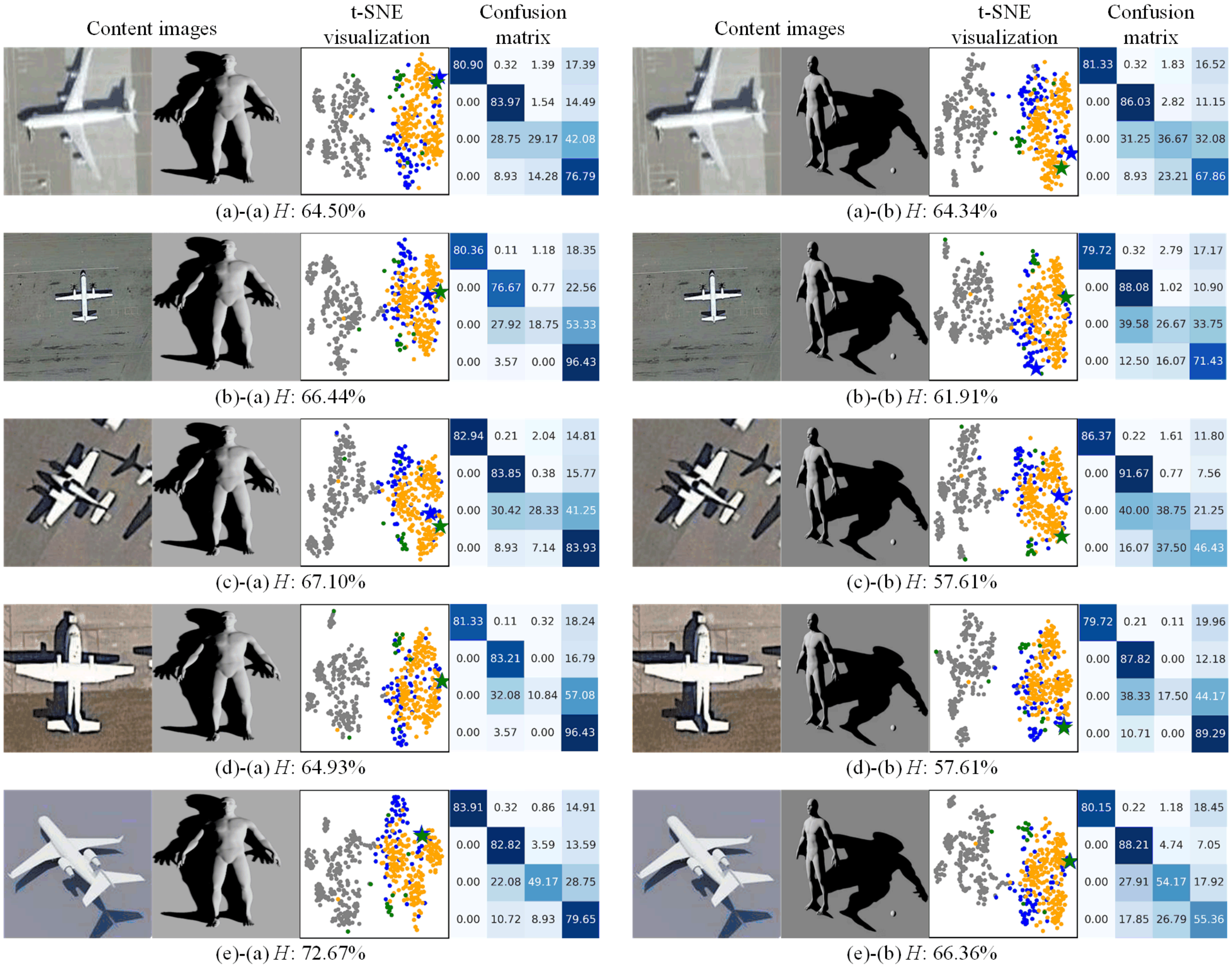
We presented instance visualizations of GZSL classification using the CCPL model enhanced with CADA, alongside comparisons to the SoftMax approach.

Fig. 6 illustrates that the SoftMax model, when trained directly, struggles with predicting the person class and often misclassifies some targets as airplanes. In contrast, while the original CCPL model improves prediction confidence for the airplane class, it still fails to accurately recognize persons. When CCPL is combined with the proposed CADA, the performance improves significantly. The enhanced model achieves high prediction confidence for seafloor, airplane, and ship classes and demonstrates a marked improvement in identifying person targets. These results align with the superior performance metrics reported in Tables III and IV, validating the effectiveness of the designed GZSSS dataset.

However, it is worth noting that some shipwrecks, which resemble frogmen, can be mistakenly identified as humans. This highlights the inherent challenges of the dataset, emphasizing the complexity of the GZSL study.

F. Feature Visualization

We visualized features extracted from the outputs of the DCNN models trained using various configurations: CCPL, CCPL with background fusion, noise modeling, and dual-loss (CCPL w/ bi&bf&n), CCPL with dual-loss and KL divergence (CCPL w/ bi&kld), and CCPL with CADA (CCPL+CADA).



(a)-(a) H: 64.50%

(a)-(b) H: 64.34%

(b)-(a) H: 66.44%

(b)-(b) H: 61.91%

(c)-(a) H: 67.10%

(c)-(b) H: 57.61%

(d)-(a) H: 64.93%

(d)-(b) H: 57.61%

(e)-(a) H: 72.67%

(e)-(b) H: 66.36%

Fig. 8. The CADA results on the CCPL generative model in the case that only one semantic content is available for each unseen class. The content image instances, t-SNE visualization, classification confusion matrices, and harmonic mean H are given. The legends in t-SNE visualization are the same in Fig. 7. In each confusion matrix, the true classes listed from top to bottom along the vertical axis are 'seabed', 'ship', 'airplane', and 'person'; the predicted classes listed from left to right along the horizontal axis are 'seabed', 'ship', 'airplane', and 'person'. The alphabetical numbers of these instances are the same in Fig. 5.

The features were extracted from both testing samples and content images and visualized using t-SNE [58], as shown in Fig. 7.

The results demonstrate that the domain augmentation with background fusion and noise modeling significantly enhances the distinction between airplanes and shipwrecks compared to the original CCPL. Contrastive KL loss without the augmentation improves the discrimination between the two unseen classes and avoid overfitting. However, the joint distribution of the SSS samples and content for the person class are cut off by the distribution from other unseen classes in the t-SNE visualization map.

Incorporating the CADA paradigm, with both domain augmentation and contrastive KL loss, further refines the discrimination among the three target classes. It effectively reduces the distribution deviation between content and SSS sample domains for the same class, enhancing the intra-class semantic differences and preventing mode collapse.

The results underline the efficacy of contrastive adaptation in distinguishing subtle semantic differences and maintaining robust feature representation across diverse domains.

G. Analysis on Each Semantic Content

We investigated the influence of different semantic contents from the unseen classes on GZSL classification performance. The analysis is illustrated in Fig. 8.

The classification performance for the person class varies depending on the source of the content. Specifically, when the person content is sourced from (a), the classification accuracy for the person class is superior compared to when the content is sourced from (b). It is due to content (a) capturing major contour characteristics of a person.

Nevertheless, the person content (a) shares certain features with the airplane class, such as the similarity between human arms and airplane wings. This contour similarity causes confusion, where persons are often misclassified as airplanes. Extreme cases can be observed in comparisons such as (b)-(a) and (d)-(a), where persons are recognized successfully but many airplanes are misunderstood as persons. Indistinguishable content distribution patterns are noted in comparisons like (b)-(b), (d)-(a), (d)-(b) and (e)-(b) due to the intra-class bias.

Interestingly, the airplane content (e), which is a shallow

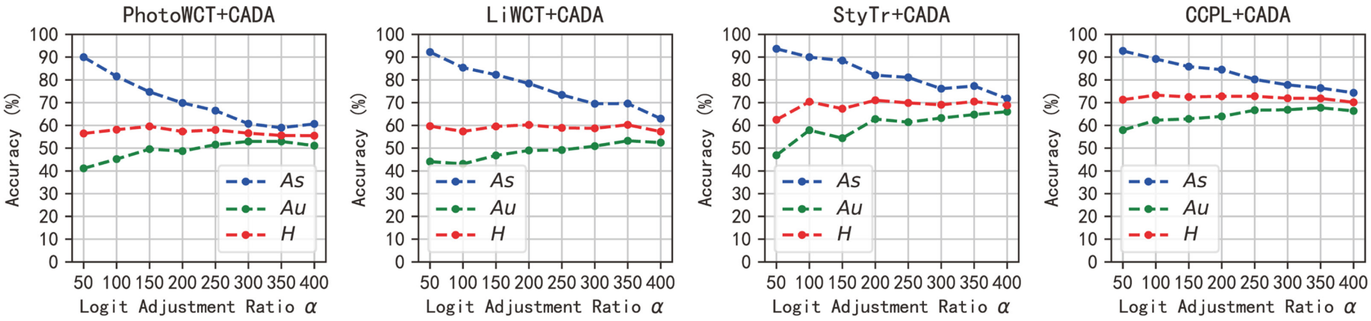


Fig. 9. Accuracy of classification under different α.

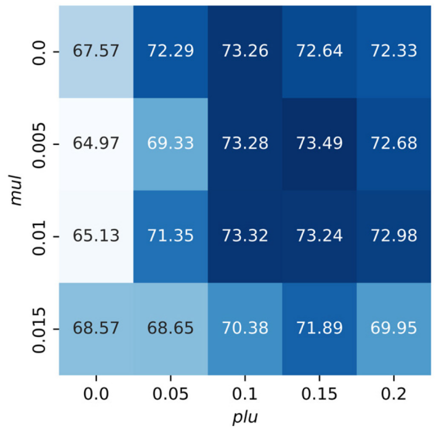


Fig. 10. The H heatmap of CCPL+CADA under different mul and plu.

photograph, delivers better results than the other four sources from RS images. This is likely because the imaging principle of (e) is closer to that of SSS and is less affected by interference.

Overall, only one semantic content achieves performance nearly as well as the default setup in our experiments. This suggests that discovering more suitable semantic photography views to mitigate domain bias is more valuable than merely increasing the number of semantic samples. Additionally, the semantic confusion between different classes remains a significant challenge.

H. Parameter Analysis

Three key hyper-parameters are involved, i.e., the logit adjustment ratio α in Eq. (12), the Rayleigh distribution scale mul and the Gaussian distribution scale plu in SSS noise addition.

We first examined α, which directly reflects the seen-unseen cognition. We explore α values ranging from 50-400, and the effects are illustrated in Fig. 9. The results show that while the classification performance for both seen and unseen classes exhibits a paradoxical relationship, an approximate optimal point can be identified. The best performance is achieved with α values of 150, 350, 200 and 100 for the PhotoWCT, LiWCT, StyTr, and CCPL with the proposed CADA. Notably, CCPL+CADA performs relatively better within the α range of 100-350. When α is set to 100, the performance aligns with the condition where the training set size is approximately 100 times the number of available semantic contents, fulfilling the mathematical requirement for class re-balancing.

Next, we consider mul and plu, which influence noise addition. The harmonic mean H results for CCPL+CADA with different scales of result of mul and plu are shown in Fig. 10. We explored mul varying from $\{0, 0.005, 0.01, 0.015, 0.02\}$ and plu from $\{0, 0.05, 0.1, 0.15\}$. The results indicate that appropriate noise addition enhances the contrastive augmentation performance. However, excessive multiplicative noise detracts from the performance. The default noise settings provide relatively optimal results.

VI. Conclusion

To address the lack of unified evaluation standards, inconsistencies in training data size, and the absence of rigorous style sample selection in underwater SSS image classification, a novel GZSL paradigm was proposed, which integrated a mature style transfer framework with background fusion and simulated SSS noise to ensure fidelity and diversity of generated samples. By applying contrastive constraints between original and generated images, the method aligned heterogeneous semantic content and enhances ability to distinguish targets.

Experimental results show that the proposed CADA achieves over a 20% improvement in the classification harmonic mean compared to both direct learning and existing generative-based methods. The incorporation of style transfer models with strong visual content preservation enhances GZSL performance, enabling accurate recognition of unseen classes under challenging conditions.

The background fusion strategy is particularly effective in preserving critical information from regions of interest while improving sample diversity, addressing challenges posed by similar grayscale distributions among unseen classes. Introducing the noise ensures model generalization and robust recognition. Moreover, higher-quality semantic photography views derived from shallow views closely resembling real target shapes can substantially improve classification accuracy while avoiding overlaps with other categories.

While contrastive adaptation reduces domain distribution gaps, intra-class bias in the unseen domain and misjudgment due to semantic likeness persist. Future work will focus on refining style transfer models to better capture semantic content, addressing residual biases, and advancing the detection of novel targets in SSS imagery.

ACKNOWLEDGMENT

This paper was supported by the National Natural Science Foundation of China (Grant No. 42276187). We would also like to thank the authors Chuanlong Li and Guanying Huo for providing the data support.

REFERENCES

[1] X. Ye, C. Li, S. Zhang, P. Yang, and X. Li, "Research on side-scan sonar image target classification method based on transfer learning," in *Proc. OCEANS 2018 MTS/IEEE Charleston*, Charleston, SC, USA, Oct. 2018, pp. 1–6.

[2] G. Huo, Z. Wu, and J. Li, "Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data," *IEEE access*, vol. 8, pp. 47407–47418, 2020, doi:10.1109/ACCESS.2020.2978880.

[3] H. Larochelle, D. Erhan, and Y. Bengio, "Zero-data learning of new tasks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 1, no. 2, 2008, p. 3.

[4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 951–958, doi:10.1109/CVPRW.2009.5206594.

[5] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2018, doi:10.1109/TPAMI.2018.2876865.

[6] Y. Wang, J. Yan, Z. Yang, Y. Wu, J. Wang, and Y. Geng, "Generative zero-shot learning for partial discharge diagnosis in gas-insulated switchgear," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023, doi:10.1109/tim.2023.3264022.

[7] D. Mahapatra, B. Bozorgtabar, and Z. Ge, "Medical image classification using generalized zero shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 3344–3353.

[8] H. K. Kim and J. Shim, "Generalized zero-shot learning for classifying unseen wafer map patterns," *Eng. Appl. Artif. Intell.*, vol. 133, Art. no. 108476, 2024.

[9] X. Wang, J. Jiao, J. Yin, W. Zhao, X. Han, and B. Sun, "Underwater sonar image classification using adaptive weights convolutional neural network," *Appl. Acoust.*, vol. 146, pp. 145–154, 2019.

[10] Y. Xu, X. Wang, K. Wang, J. Shi, and W. Sun, "Underwater sonar image classification using generative adversarial network and convolutional neural network," *IET Image Process.*, vol. 14, no. 12, pp. 2819–2825, 2020, doi:10.1049/ipr2.12406.

[11] C. Li, X. Ye, D. Cao, J. Hou, and H. Yang, "Zero shot objects classification method of side scan sonar image based on synthesis of pseudo samples," *Appl. Acoust.*, vol. 173, Art. no. 107691, 2021.

[12] H. Xu, Z. Bai, X. Zhang, and Q. Ding, "Mfsanet: Zero-shot side-scan sonar image recognition based on style transfer," *IEEE Geosci. Remote Sens. Lett.*, 2023, doi: 10.1109/LGRS.2023.3318051.

[13] Z. Bai, H. Xu, Q. Ding, and X. Zhang, "Side-scan sonar image classification with zero-shot and style transfer," *IEEE Trans. Instrum. Meas.*, 2024.

[14] C. Huang, J. Zhao, Y. Yu, and H. Zhang, "Comprehensive sample augmentation by fully considering SSS imaging mechanism and environment for shipwreck detection under zero real samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[15] W. Jiao and J. Zhang, "Sonar images classification while facing long-tail and few-shot," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2022, doi:10.1109/TGRS.2022.3202843.

[16] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1501–1510, doi:10.1109/ICCV.2017.162.

[17] X. Li, S. Liu, J. Kautz, and M.-H. Yang, "Learning linear transformations for fast image and video style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3809–3817, doi:10.1109/cvpr.2019.00393.

[18] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "StyTR2: Image style transfer with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11326–11336.

[19] Z. Wu, Z. Zhu, J. Du, and X. Bai, "CCPL: Contrastive coherence preserving loss for versatile style transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel, Oct. 2022, pp. 189–206, doi:10.1007/978-3-031-19947-4_12.

[20] J. Chen and J. E. Summers, "Deep convolutional neural networks for semi-supervised learning from synthetic aperture sonar (SAS) images," in *Proc. Meet. Acoust.*, vol. 30, no. 1, New Orleans, LA, USA, Jun. 2017, Art. no. 055002, doi:10.1121/2.0001018.

[21] Y. Steiniger, D. Kraus, and T. Meisen, "Survey on deep learning based computer vision for sonar imagery," *Eng. Appl. Artif. Intell.*, vol. 114, Art. no. 105157, 2022, doi:10.1016/j.engappai.2022.105157.

[22] P. Zhang, J. Tang, H. Zhong, M. Ning, D. Liu, and K. Wu, "Self-trained target detection of radar and sonar images using automatic deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021, doi:10.1109/TGRS.2021.3056957.

[23] T. Berthold, A. Leichter, B. Rosenhahn, V. Berkhahn, and J. Valerius, "Seabed sediment classification of side-scan sonar data using convolutional neural networks," in *2017 IEEE Symp. Series Comput. Intell. (SSCI)*, Honolulu, HI, USA, Nov. 2017, pp. 1–8, doi:10.1109/ssci.2017.8285220.

[24] X. Luo, X. Qin, Z. Wu, F. Yang, M. Wang, and J. Shang, "Sediment classification of small-size seabed acoustic images using convolutional neural networks," *IEEE Access*, vol. 7, pp. 98331–98339, 2019, doi:10.1109/ACCESS.2019.2927051.

[25] T. T. Chungath, A. M. Nambiar, and A. Mittal, "Transfer learning and few-shot learning based deep neural network models for underwater sonar image classification with a few samples," *IEEE J. Oceanic Eng.*, vol. 49, no. 1, pp. 294–310, 2023, doi:10.1109/JOE.2022.3221127.

[26] Z. Dai, H. Liang, and T. Duan, "Small-sample sonar image classification based on deep learning," *J. Mar. Sci. Eng.*, vol. 10, no. 12, p. 1820, 2022, doi:10.3390/jmse10121820.

[27] J. Xi and X. Ye, "Sonar image target detection based on simulated stain-like noise and shadow enhancement in optical images under zero-shot learning," *J. Mar. Sci. Eng.*, vol. 12, no. 2, p. 352, 2024.

[28] X. Qin, X. Luo, Z. Wu, and J. Shang, "Optimizing the sediment classification of small side-scan sonar images based on deep learning," *IEEE Access*, vol. 9, pp. 29416–29428, 2021, doi:10.1109/ACCESS.2021.3052206.

[29] J. Wang, H. Li, G. Huo, C. Li, and Y. Wei, "Multi-modal multi-stage underwater side-scan sonar target recognition based on synthetic images," *Remote Sensing*, vol. 15, no. 5, p. 1303, 2023.

[30] D. Yang, C. Wang, C. Cheng, G. Pan, and F. Zhang, "Data generation with GAN networks for sidescan sonar in semantic segmentation applications," *J. Mar. Sci. Eng.*, vol. 11, no. 9, p. 1792, 2023.

[31] Z. Xi, J. Zhao, and W. Zhu, "Side-scan sonar image simulation considering imaging mechanism and marine environment for zero-shot shipwreck detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023, doi:10.1109/TGRS.2023.3247487.

[32] Y. Jia, X. Ye, Y. Liu, and S. Guo, "Multi-modal recursive prompt learning with mixup embedding for generalization recognition," *Knowl.-Based Syst.*, vol. 294, Art. no. 111726, 2024, doi:10.1016/j.knosys.2024.111726.

[33] W. Jiao, J. Zhang, and C. Zhang, "Open-set recognition with long-tail sonar images," *Expert Syst. Appl.*, vol. 249, Art. no. 123495, 2024, doi:10.1016/j.eswa.2024.123495.

[34] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," *Advances Neural Inf. Process. Syst.*, vol. 22, 2009.

[35] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5652–5667, 2018, doi:10.1109/TIP.2018.2840874.

[36] V. K. Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4281–4289.

[37] R. Keshari, R. Singh, and M. Vatsa, "Generalized zero-shot learning via over-complete distribution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13300–13308, doi:10.1109/CVPR42600.2020.01331.

[38] V. K. Verma, D. Brahma, and P. Rai, "Meta-learning for generalized zero-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 04, 2020, pp. 6062–6069, doi:10.1609/AAAI.V34I04.6069.

[39] D. Chen, Y. Shen, H. Zhang, and P. H. Torr, "Zero-shot logit adjustment," *arXiv preprint arXiv:2204.11822*, 2022.

[40] S. Chen, Z. Hong, W. Hou, G.-S. Xie, Y. Song, J. Zhao, X. You, S. Yan, and L. Shao, "Transzero++: Cross attribute-guided transformer for zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12844–12861, 2022, doi:10.1109/TPAMI.2022.3229526.

[41] R. Gao, X. Hou, J. Qin, J. Chen, L. Liu, F. Zhu, Z. Zhang, and L. Shao, "Zero-vae-gan: Generating unseen features for generalized and transductive zero-shot learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3665–3680, 2020, doi:10.1109/TIP.2020.2966506.

[42] Z. Chen, Y. Luo, R. Qiu, S. Wang, Z. Huang, J. Li, and Z. Zhang, "Semantics disentangling for generalized zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8712–8720.

[43] Z. Han, Z. Fu, S. Chen, and J. Yang, "Semantic contrastive embedding for generalized zero-shot learning," *Int. J. Comput. Vis.*, vol. 130, no. 11, pp. 2606–2622, 2022, doi:10.1007/s11263-022-01656-y.

[44] Y. Du, M. Shi, F. Wei, and G. Li, "Boosting zero-shot learning via contrastive optimization of attribute representations," *IEEE Trans. Neural Networks Learn. Syst.*, 2023.

[45] Y. Wang, J. Mao, C. Guo, and S. Chen, "Contrastive prototype-guided generation for generalized zero-shot learning," *Neural Networks*, vol. 176, Art. no. 106324, 2024, doi:10.1016/j.neunet.2023.106324.

[46] Z. Ji, H. Wang, Y. Yu, and Y. Pang, "A decadal survey of zero-shot image classification," *Scientia Sinica: Inf.*, vol. 49, no. 10, pp. 1299–1320, 2019, doi:10.1360/N112018-00312.

[47] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th Eur. Conf. Comput. Vis., Amsterdam, The Netherlands*, Oct. 11-14, 2016, Proc., Part II, vol. 14, Springer, 2016, pp. 694–711.

[48] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," *Advances in Neural Inf. Process. Syst.*, vol. 30, 2017.

[49] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz, "A closed-form solution to photorealistic image stylization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 453–468, doi:10.1007/978-3-030-01219-9_28.

[50] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning," *J. Big Data*, vol. 4, pp. 1–42, 2017, doi:10.1186/s40537-017-0089-0.

[51] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *Proc. AAAI Conf. Artif. Intell.*, vol. 3, 2006, p. 5.

[52] Y. Jia, X. Ye, S. Guo, and H. Yang, "A piecewise nonlinear enhancement method of side scan sonar images," in *OCEANS 2019-Marseille, Marseille*, France, 2019, pp. 1–6, doi:10.1109/OCEANSE.2019.8867371.

[53] C. Peng, S. Fan, X. Cheng, Y. Cao, and G. Zeng, "An improved side scan sonar image processing framework for autonomous underwater vehicle navigation," in *Proc. 15th Int. Conf. Underwater Networks Syst.*, 2021, pp. 1–5.

[54] Y. Liu and X. Ye, "A gray scale correction method for side-scan sonar images considering rugged seafloor," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–10, 2023.

[55] M. Sung, Y.-W. Song, and S.-C. Yu, "Underwater object detection of auv based on sonar simulator utilizing noise addition," in *2022 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*, Monterey, CA, USA, 2022, pp. 1–5.

[56] C. S. Seelamantula and T. Blu, "Image denoising in multiplicative noise," in *2015 IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, 2015, pp. 1528–1532.

[57] M. K. Mihcak, I. Kozintsev, K. Ramchandran, and P. Moulin, "Low-complexity image denoising based on statistical modeling of wavelet coefficients," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 300–303, 1999.

[58] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008, doi:10.1162/jmlr.2008.9.11.2579.

Yunpeng Jia received the B.E. degree in College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China, in 2015, and is currently working toward Ph.D. degree in College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin, China. His research interests include image processing, classification and zero-shot learning.

Xiufen Ye (Senior Member, IEEE), received the B.S., M.S., and Ph.D. degrees in control theory and control engineering from Harbin Shipbuilding Engineering University (Harbin Engineering University), Harbin, China, in 1987, 1990, and 2003, respectively.

She is currently a Professor with Harbin Engineering University. Her research interests include underwater vehicle intelligent control systems, digital image processing, and object detection and tracking. She has served as the Program Committee Chairs for the IEEE ICIA 2010, the IEEE/ICME CME 2011, and the IEEE ICMA 2023.

Peng Li is an associate professor in Harbin University of Commerce. He received his PhD degree in the school of automation in Harbin Engineering University in 2016. He received his B.E. degree in Automation from Harbin University of Science and Technology in 2008 and M.E. degree in Pattern Recognition and Intelligence System from Harbin Engineering University in 2012. He finished his postdoctoral research of recommender system in Harbin University of Commerce. His research interests include recommender system, computer vision, and machine learning.

Shuxiang Guo (Fellow, IEEE) received the Ph.D. degree in mechano-informatics and systems from Nagoya University, Nagoya, Japan, in 1995. He is currently a Chair Professor with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen, Guangdong, China. He is also a Chair Professor with the Key Laboratory of Convergence System and Healthcare Technology Medical Engineering, The Ministry of Industry and Information Technology, Beijing Institute of Technology, Beijing, China. His research interests include medical robot systems, microcatheter systems, and biomimetic underwater robots. Dr. Guo has a fellowship of The Engineering Academy of Japan.