



## Article

# Generalized Zero-Shot Space Target Recognition Based on Global-Local Visual Feature Embedding Network

Yuanpeng Zhang <sup>1,2,†</sup> , Jingye Guan <sup>3,†</sup> , Haobo Wang <sup>2</sup> , Kaiming Li <sup>2,\*</sup> , Ying Luo <sup>2</sup> and Qun Zhang <sup>2,4</sup><sup>1</sup> Early Warning Academy, Wuhan 430019, China; zhangyuanpeng312@163.com<sup>2</sup> Information and Navigation College, Air Force Engineering University, Xi'an 710077, China; whbwhnkgd@163.com (H.W.); luoying2002521@163.com (Y.L.); afeuzq@163.com (Q.Z.)<sup>3</sup> Sichuan Sen Yu Hu Lian Technology Co., Ltd., Chengdu 610000, China; guanjymail@126.com<sup>4</sup> Key Laboratory for Information Science of Electromagnetic Waves, Fudan University, Shanghai 200433, China

\* Correspondence: likaiming1982@163.com

† These authors contributed equally to this work.

**Abstract:** Existing deep learning-based space target recognition methods rely on abundantly labeled samples and are not capable of recognizing samples from unseen classes without training. In this article, based on generalized zero-shot learning (GZSL), we propose a space target recognition framework to simultaneously recognize space targets from both seen and unseen classes. First, we defined semantic attributes to describe the characteristics of different categories of space targets. Second, we constructed a dual-branch neural network, termed the global-local visual feature embedding network (GLVFENet), which jointly learns global and local visual features to obtain discriminative feature representations, thereby achieving GZSL for space targets with higher accuracy. Specifically, the global visual feature embedding subnetwork (GVFE-Subnet) calculates the compatibility score by measuring the cosine similarity between the projection of global visual features in the semantic space and various semantic vectors, thereby obtaining global visual embeddings. The local visual feature embedding subnetwork (LVFE-Subnet) introduces soft space attention, and an encoder discovers the semantic-guided local regions in the image to then generate local visual embeddings. Finally, the visual embeddings from both branches were combined and matched with semantics. The calibrated stacking method is introduced to achieve GZSL recognition of space targets. Extensive experiments were conducted on an electromagnetic simulation dataset of nine categories of space targets, and the effectiveness of our GLVFENet is confirmed.

**Keywords:** radar recognition; space targets; high-resolution range profile (HRRP) sequence; micromotion; zero-shot learning (ZSL)



**Citation:** Zhang, Y.; Guan, J.; Wang, H.; Li, K.; Luo, Y.; Zhang, Q. Generalized Zero-Shot Space Target Recognition Based on Global-Local Visual Feature Embedding Network. *Remote Sens.* **2023**, *15*, 5156. <https://doi.org/10.3390/rs15215156>

Academic Editors: Yanhua Wang, Liang Zhang, Shunqiao Sun and Pu Wang

Received: 17 September 2023

Revised: 23 October 2023

Accepted: 25 October 2023

Published: 28 October 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the increase in human space exploration, the number of micromotion space targets, such as satellites, warheads, and space debris, is increasing. Radar is the main sensor used for space target detection, tracking, and recognition. In the radar automatic target recognition field, extracting target information (such as shape, structure, and micromotion characteristics) from radar echoes for space target recognition has attracted widespread attention [1–4].

The existing methods for space target recognition can be categorized as model-driven methods [2,5,6] and data-driven methods [7–9]. The model-driven methods require hand-crafted features, leading to high complexity and weak generalizability [10]. The data-driven methods based on deep learning do not require manually designed features and can extract discriminative target features from the data automatically. Researchers have begun using deep learning methods to recognize space targets, achieving higher recognition accuracy than model-driven methods. Notably, these deep learning methods can recognize only targets belonging to the seen classes during training. When unseen classes of space targets

appear during the testing phase, these methods become ineffective, thereby limiting the practical applications of the algorithms. In the real world, encountering targets during the testing phase that are not seen during the training phase is likely, especially for non-cooperative targets whose samples cannot be obtained in advance. Therefore, methods for recognizing unseen classes that appear during the testing phase must be studied.

Inspired by human learning of new concepts, zero-shot learning (ZSL) techniques have attracted increasing interest [11]. Deep learning-based supervised classification methods can recognize only targets belonging to the seen classes during the training phase. ZSL recognizes unseen classes in the testing set by transferring semantic knowledge from seen to unseen classes, where there are no available training samples. To our knowledge, no space target recognition methods based on ZSL currently exist.

In the computer vision field, existing ZSL methods for target recognition can be categorized into three main types: early embedding-based methods [12–15], generative-based methods [16–19], and part embedding-based methods [20–23]. Early embedding-based methods learn mapping spaces between the visual features of seen classes and their semantic descriptions. They recognize unseen classes by mapping them to these spaces and conducting nearest-neighbor searches. One major issue with these methods is the potential bias problem [24]. To alleviate this issue, generative-based methods utilize generative adversarial networks (GANs) to generate features for unseen classes, thereby transforming ZSL into fully supervised classification problems. Although these methods have exhibited certain improvements in unseen class recognition accuracy, they consider only global visual features and cannot capture the local regions in images that correspond to semantic descriptions, limiting the transfer of semantic knowledge. Recently, some part embedding-based methods have achieved higher recognition accuracy by incorporating attention mechanisms or obtaining discriminative visual features from local regions in images guided by semantic attributes. However, we believe that both global and local visual features play important roles in ZSL. Additionally, due to the lack of semantic attributes specifically defined for space targets, it is difficult to directly apply existing methods to the ZSL of space targets. Therefore, we attempt to design binary semantic attributes for space targets and propose a novel ZSL method that simultaneously considers global and local visual features, thereby improving the ZSL capability for space targets.

In this article, an end-to-end framework for generalized zero-shot space target recognition, termed the Global-Local Visual Feature Embedding Network (GLVFENet), is proposed. This framework is used to simultaneously recognize both seen and unseen space targets during the testing phase. First, we devise prior binary semantic attributes for each space target category, which is the key to ZSL. Next, we introduce a CNN backbone network that extracts initial features from raw high-range resolution profile (HRRP) sequences. The output of this network is then passed to two subsequent subnetworks. Then, we develop a global visual feature embedding subnetwork (GVFE-Subnet) that can further capture global visual features from the output of the CNN backbone network and map these features to the semantic space. By using cosine similarity to calculate the compatibility scores with the ground-truth semantic attributes, we obtain the global visual embeddings. Because the discriminative visual properties in the images are the key to transferring knowledge from seen classes to unseen classes, we designed a local visual feature embedding subnetwork (LVFE-Subnet). LVFE-Subnet utilizes soft space attention and an encoder to localize discriminative local regions in the images through semantic knowledge, obtaining local visual embeddings. Finally, we optimize the full network by using a joint loss function and consider both global and local visual embeddings to achieve GZSL recognition of space targets. The following are the contributions of this article:

- (1) A framework for GZSL of space targets is proposed for recognizing targets of both seen and unseen classes. The framework consists primarily of a GVFE-Subnet and an LVFE-Subnet, and the dual-branch network improves the model's ability to transfer knowledge from seen classes to unseen classes.

(2) To our knowledge, this is the first attempt to utilize ZSL techniques for space target recognition. To achieve ZSL, we incorporate expert prior knowledge in the space target field and design binary semantic attributes to describe the characteristics of different categories of space targets, forming the basis for inferring unseen classes.

(3) The results of comparative and ablation experiments based on electromagnetic (EM) calculation data demonstrate the effectiveness of the proposed method.

The remainder of the article is organized as follows: Section 2 introduces the related work on space target recognition and ZSL. Section 3 provides a detailed description of the proposed GLVFENet. Section 4 describes the dataset generation, evaluation metrics, and implementation details and presents the experimental results. Section 5 includes discussions. Finally, Section 6 concludes the article.

## 2. Related Work

### 2.1. Space Target Recognition

Space target recognition methods can be divided into model-driven methods and data-driven methods. Model-driven methods include feature extraction and classifier design. The extracted features can be further classified into physically meaningful features [2,25–28] and features without explicit physical meanings [3,4,29–31]. After feature extraction, the extracted target features are fed into a classifier to achieve recognition. Commonly used classifiers include KNN classifiers [3,30] and SVM classifiers [29]. Model-driven methods rely on manual features designed by experts, resulting in high complexity and weak generalization performance [10]. Data-driven methods can automatically extract target features from a large amount of data without hand-crafted features, improving recognition accuracy, and the classifier and feature extractor are integrated within a unified framework. The methods can be roughly divided into convolutional neural network (CNN)-based methods [7,8,32–39], recurrent neural network (RNN)-based methods [9,40], and encoder-decoder-based methods [41,42] based on the network used. However, existing data-driven methods can recognize only the classes that have already been seen in the test set, and their recognition performance sharply decreases for unseen classes. To this end, we propose a ZSL method for space targets that can simultaneously recognize seen and unseen classes in the test set.

### 2.2. Zero-Shot Learning

ZSL methods can be categorized into conventional zero-shot learning (CZSL) methods and generalized zero-shot learning (GZSL) methods based on whether the seen classes are visible during the testing phase. CZSL methods predict only the unseen classes in the test set, while GZSL methods predict both unseen and seen classes in the test set. GZSL methods are more practical and challenging. In this paper, a space target recognition method is proposed under the GZSL framework. ZSL methods can be further categorized into early embedding-based, generative-based, and part-based methods. Early embedding-based methods focus primarily on learning embedding spaces where visual features and semantic knowledge can interact. In this embedding space, the global visual features extracted from images and semantic descriptions (such as manually annotated attributes [43], word vectors learned through language models [44], and textual descriptions [45]) are aligned. Since unseen and seen classes share the same embedding spaces, semantic knowledge aligned with visual representations can be transferred to unseen classes for recognition. Based on this paradigm, [12,13,46] map global visual features to semantic spaces, and [14] maps semantic knowledge to visual spaces. Refs. [15,47] map visual features and semantic knowledge to the same spaces. One problem with these methods is their bias toward classifying seen classes as unseen classes, known as the bias problem. Generative-based methods have been proposed to address this issue. These methods utilize generative models to generate features for unseen classes, transforming ZSL into fully supervised classification problems. Currently, the main generative models used for zero-shot classification tasks include generative adversarial networks (GANs) [16,17] and variational autoencoders

(VAEs) [18,19]. However, most of these methods consider only global visual features, which are prone to interference from background noise. Additionally, global features of similar classes can also be similar, hindering the transfer of semantic knowledge. Recent part-based methods have been able to capture discriminative features in images by localizing specific regions. This has resulted in better alignment between visual features and semantic knowledge, enhancing the ability of these methods to transfer semantic knowledge from seen classes to unseen classes. One approach to these methods is utilizing attention mechanisms to discover local regions in the images [20,21]. Another approach is localizing the corresponding local regions in images guided by the semantic knowledge [22,23]. Inspired by this, we simultaneously utilize both global visual representations and locally discriminative visual features localized by semantic descriptions in the images. This approach aims to better align visual features and semantic knowledge, thereby improving the transferability of semantic knowledge from seen classes to unseen classes and enhancing the recognition of both seen and unseen classes.

### 3. Proposed Method

In this section, a formulaic definition of the task is first provided, then the semantic information proposed for space target recognition is introduced, and finally a detailed description of the proposed method is presented.

#### 3.1. Problem Formulation

Let  $S = \left\{ (\mathbf{X}_i^s, y_i^s, \mathbf{a}_i^s)_{i=1}^{N_s} \mid \mathbf{X}_i^s \in \mathcal{X}^s, y_i^s \in \mathcal{Y}^s, \mathbf{a}_i^s \in \mathcal{A}^s \right\}$  be the seen class set, where  $\mathbf{X}_i^s \in \mathbb{R}^{H \times Q}$  denotes an HRRP sequence with duration  $H$  and the number of single HRRP range cells  $Q$ ,  $y_i^s$  denotes the label corresponding to  $\mathbf{X}_i^s$ ,  $\mathcal{Y}^s = \{l_1^s, l_2^s, \dots, l_{C_s}^s\}$  denotes the set of labels of the seen class,  $C_s$  denotes the number of seen classes,  $\varphi(y_i^s) \in \{0, 1\}^K$  denotes a semantic vector of seen classes with  $K$  dimensions, and  $N_s$  denotes the number of samples of seen classes. Similarly, the dataset of unseen classes can be denoted as  $\mathcal{U} = \left\{ (\mathbf{X}_i^u, y_i^u, \mathbf{a}_i^u)_{i=1}^{N_u} \mid \mathbf{X}_i^u \in \mathcal{X}^u, y_i^u \in \mathcal{Y}^u, \mathbf{a}_i^u \in \mathcal{A}^u \right\}$ , where the HRRP sequence  $\mathbf{X}_i^u \in \mathbb{R}^{H \times Q}$  is available only in the testing phase,  $y_i^u$  denotes the labels corresponding to  $\mathbf{X}_i^u$ ,  $\mathcal{Y}^u = \{l_1^u, l_2^u, \dots, l_{C_u}^u\}$  denotes the set of labels of the unseen classes,  $C_u$  denotes the number of unseen classes, and  $\varphi(y_i^u) \in \{0, 1\}^K$  denotes an unseen class semantic vector.  $\mathcal{X} = \mathcal{X}^s \cup \mathcal{X}^u$  denotes the sample space consisting of all seen and unseen classes,  $\mathcal{Y} = \mathcal{Y}^s \cup \mathcal{Y}^u$  denotes the label space, and  $\mathcal{Y}^s \cap \mathcal{Y}^u = \emptyset$ ,  $\mathcal{A} = \mathcal{A}^s \cup \mathcal{A}^u$  denotes the semantic space. In the GZSL task, all the training classes come from  $S$ , and the test dataset can be denoted as  $\mathcal{T} = \left\{ (\mathbf{X}_m^{ts}, y_m^{ts})_{m=1}^{N_t} \mid \mathbf{X}_m^{ts} \in \mathcal{X}, y_m^{ts} \in \mathcal{Y} \right\}$ , where  $N_t$  denotes the number of samples in the test set. The goal is to train a model  $f: \mathcal{X} \rightarrow \mathcal{Y}$  to recognize  $N_t$  samples from  $\mathcal{T}$ .

#### 3.2. Semantic Representation

The semantic information is crucial for ZSL. Since samples from unseen classes are unavailable during the training phase, establishing a relationship between seen and unseen classes through semantic information is necessary. However, to our knowledge, no semantic information specifically designed for describing space targets currently exists. In this section, we attempt to design semantic information for the ZSL of space targets based on the HRRP sequence. The designed semantic information consists of multiple binary semantic attributes and follows the following design principles: (1) Shareability: Each attribute can be shared by multiple categories, and the semantic space formed by all attributes can be shared by both seen and unseen classes. (2) Distinguishability: The semantics of each category, formed by attributes, must be distinct, providing accurate guidance for the network. (3) Interpretability: The designed attributes should reflect discriminative visual features and be interpretable or understandable.

Following the semantic design principles, we define binary semantic attributes for space targets. As the HRRP sequence contains both geometric and micro-Doppler char-

acteristics of the targets, we mainly focus on these two aspects when defining the binary semantic attributes. Specifically, we define six obvious appearance attributes, including two trajectories, three trajectories, four trajectories, sine curve, amplitude size, and coupling degree. These attributes can be directly observed from the HRRP sequences. However, they are not sufficient for distinguishing all nine categories of targets. We thus also add three derived semantic attributes, including double rotational symmetry, nutation, and wobble. Although these attributes cannot be directly obtained from the images, they help distinguish the targets. Finally, we group these nine binary semantic attributes into geometric features and micro-Doppler features, as shown in Table 1. The value “1” indicates that the class of targets has this attribute, while “0” indicates that the class of targets does not have this attribute.

**Table 1.** Binary attributes based on the HRRP sequence of the space target.

Target Type	Geometric Features					Micromotion Features			
	1	2	3	4	5	6	7	8	9
T1	1	0	0	0	1	0	0	0	0
T2	1	0	0	0	0	0	0	1	0
T3	1	0	0	0	0	0	0	0	1
T4	1	1	0	0	1	1	1	0	0
T5	1	1	0	1	0	1	1	0	0
T6	0	1	0	0	1	0	0	0	0
T7	0	1	0	0	0	0	0	1	0
T8	0	1	0	0	0	0	0	0	1
T9	0	1	1	0	1	1	1	0	0

The numbers 1–9 represent 9 different attributes, which are two trajectories, three trajectories, four trajectories, double rotational symmetry, sine curve, amplitude size, coupling degree, nutation, and wobble, respectively.

### 3.3. GLVFENet

As shown in Figure 1, GLVFENet consists of three parts: a shared feature extraction module, GVFE-Subnet, and LVFE-Subnet. The shared feature extraction module is used to extract initial features from the raw HRRP sequences, and its outputs are then fed into the two subsequent subnetworks. The GVFE-Subnet uses cosine similarity to calculate the compatibility scores between the projection of global visual features in the semantic space and the semantic vectors of each class. Under the supervision of the loss function, the GVFE-Subnet learns the global visual embeddings in the image. Since the visual features learned by the GVFE-Subnet are coarse-grained, distinguishing between seen and unseen classes with subtle differences in global feature spaces is difficult. Therefore, we design an LVFE-Subnet. This subnetwork utilizes soft space attention and an encoder to obtain semantically guided local visual embeddings. Finally, we summarize the end-to-end overall loss function and provide a prediction method for sample labels.

#### 3.3.1. Shared Feature Extraction Module

CNN is a commonly used backbone network in visual tasks [48] that is known for its strong feature representation ability. To this end, a CNN is used as a feature extraction network at the beginning of the network to capture features from the raw HRRP sequence. Let the  $i$ -th sample of the HRRP sequence fed into the network be denoted as  $\mathbf{X}_i \in \mathbb{R}^{H_0 \times Q_0}$ , where  $H_0$  denotes the duration of the HRRP sequence and  $Q_0$  denotes the number of range cells of a single HRRP. The HRRP sequence representation obtained after CNN operation is denoted as  $\mathbf{X}_i^{\text{CNN}} = \text{Conv}(\mathbf{X}_i) \in \mathbb{R}^{C \times H \times Q}$ , where  $C$ ,  $H$  and  $Q$  denote the number of channels, height, and width of the feature map, respectively, and  $\text{Conv}(\cdot)$  denotes the convolution operation. Next,  $\mathbf{X}_i^{\text{CNN}}$  is fed into GVFE-Subnet and LVFE-Subnet.

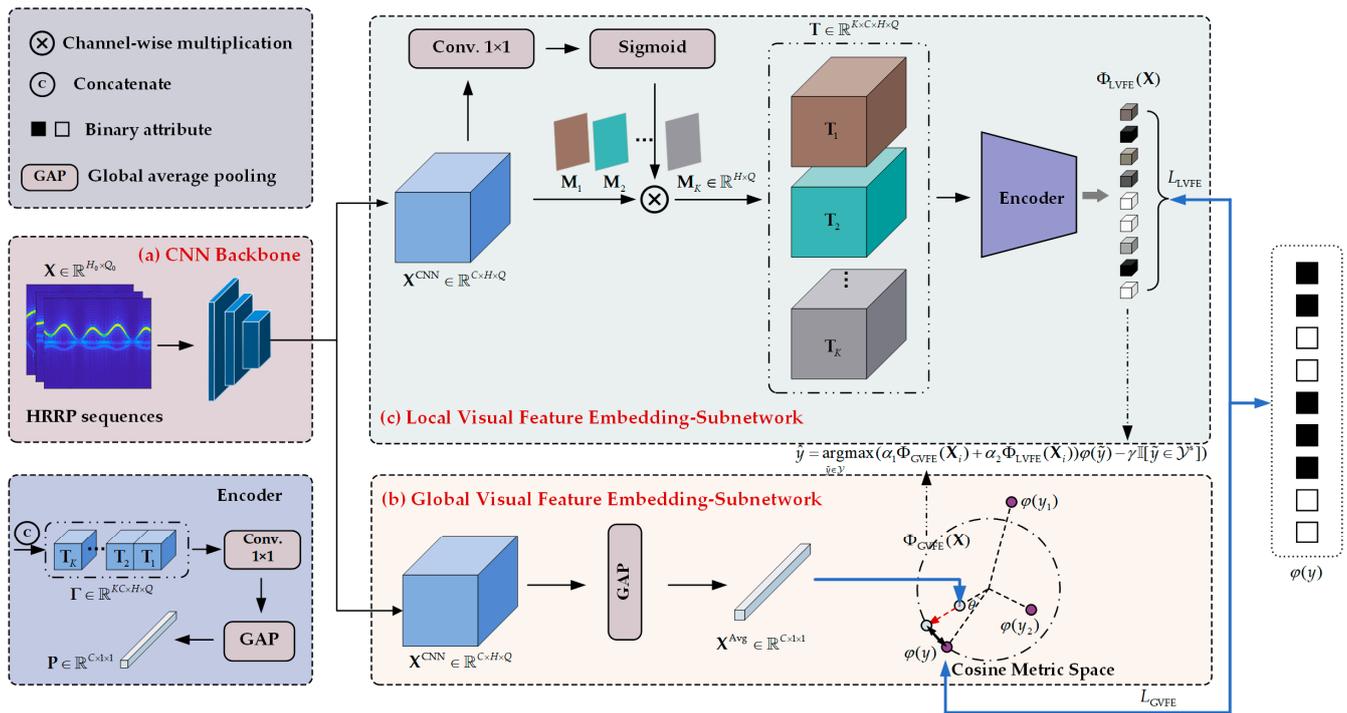


Figure 1. Overall structure of GLVFENet.

### 3.3.2. Global Visual Feature Embedding Subnet

The global visual feature embedding subnet (GVFE-Subnet) is employed to further capture global visual features from HRRP sequence representations and map these features to semantic spaces. Then, in the cosine metric space, the compatibility scores between visual feature embedding and semantic vectors are calculated, and coarse-grained global visual embeddings are obtained under the supervision of the loss function.

First, a global average pooling operation is performed on the input  $X_i^{CNN}$  of this subnet. The result  $X_i^{Avg} \in \mathbb{R}^{C \times 1 \times 1}$  can be expressed as follows:

$$X_i^{Avg} = \text{Avgpooling}(X_i^{CNN}) \tag{1}$$

where  $\text{Avgpooling}(\cdot)$  represents the average pooling operation.

The compatibility-based approach is widely used in ZSL tasks. In this approach, a compatibility function that measures the degree of compatibility between the image representation and the semantic vector of a sample is learned. The sample is labeled as the class corresponding to the semantic vector with the highest compatibility score with the image representation. Specifically, a learnable weight  $V \in \mathbb{R}^{C \times K}$  is first applied to  $X_i^{Avg}$  to project global visual features into the semantic space.

$$\Phi_{GVFE}(X_i) = (X_i^{Avg})^T V \tag{2}$$

Then, the compatibility score between the visual projection  $\Phi_{GVFE}(X_i)$  and each semantic vector is computed as follows:

$$C(X_i^{Avg}, \varphi(y_j^s)) = \Phi_{GVFE}(X_i) \times \varphi(y_j^s), j = 1, \dots, C_s + C_u \tag{3}$$

where  $C(\cdot, \cdot)$  represents the compatibility function.

Cosine similarity is commonly used to calculate the similarity of text embeddings, which helps limit and reduce the variance of neurons, resulting in models with better generalizability [49]. We use cosine similarity to predict which class  $X_i^{Avg}$  belongs to. In this

way, the output of cosine similarity is the cosine value of the compatibility score between visual embedding  $\mathbf{X}_i^{\text{Avg}} \cdot \mathbf{V}$  and each semantic vector  $\varphi(y_j^s)$ . Therefore, the probability of label  $y_j$  for sample  $\mathbf{X}_i^{\text{Avg}}$  can be expressed as follows:

$$P(y_i = y_j | \mathbf{X}_i^{\text{Avg}}) = \frac{\exp(\sigma \cos(C(\mathbf{X}_i^{\text{Avg}}, \varphi(y_j))))}{\sum_{l \in \mathcal{Y}} \exp(\sigma \cos(C(\mathbf{X}_i^{\text{Avg}}, \varphi(y_l))))} \quad (4)$$

where  $\sigma$  represents the scaling factor.

Finally, a loss function based on cross-entropy is used to encourage the visual projection of input samples to have the highest compatibility score with their corresponding semantic vectors. This function is defined as follows:

$$L_{\text{GVFE}} = -\frac{1}{N_b} \sum_{j=1}^{N_b} y_j \log \frac{\exp(\sigma \cos(C(\mathbf{X}_i^{\text{Avg}}, \varphi(y_j))))}{\sum_{l \in \mathcal{Y}} \exp(\sigma \cos(C(\mathbf{X}_i^{\text{Avg}}, \varphi(y_l))))} \quad (5)$$

where  $N_b$  represents the number of samples in a mini-batch.

### 3.3.3. Local Visual Feature Embedding Subnet

The GVFE-Subnet captures the overall global visual features of the targets. However, when subtle differences in the global feature space exist between seen and unseen classes, relying solely on the global space may make distinguishing these subtle differences difficult. The local regions in an image can be abstracted into semantic attributes at high network levels, and the local regions corresponding to semantic attributes have stronger discriminative ability [50,51]. Therefore, we designed a local visual feature embedding subnet (LVFE-Subnet) based on a soft spatial attention [20] and a visual-semantic embedding encoder to obtain fine-grained local visual embeddings in the image guided by semantics.

(1) *Soft spatial attention*: Soft spatial attention is used to map the output  $\mathbf{X}_i^{\text{CNN}}$  of the shared feature extraction module into  $K$  attention feature maps. A  $1 \times 1$  convolution is first used to perform a convolution operation with  $\mathbf{X}_i^{\text{CNN}}$  and generate  $K$  attention mask maps  $\mathbf{M} \in \mathbb{R}^{K \times H \times Q}$  via a sigmoid activation function:

$$\mathbf{M} = \text{Sigmoid}(\text{Conv}_{1 \times 1}(\mathbf{X}_i^{\text{CNN}})) \quad (6)$$

where  $\text{Conv}_{1 \times 1}(\cdot)$  and  $\text{Sigmoid}(\cdot)$  represent the  $1 \times 1$  convolution and sigmoid activation function operations, respectively, and  $\mathbf{M}_i \in \mathbb{R}^{H \times Q}$  can be sliced from  $\mathbf{M} = \{\mathbf{M}_i\}_{i=1}^K$ .

Then, we extend  $\mathbf{M}_i$  to the same dimensions as  $\mathbf{X}_i^{\text{CNN}}$  and perform the corresponding channel-by-element multiplication operation with  $\mathbf{M}_i$  to obtain the  $K$  attention feature maps  $\mathbf{T} = \{\mathbf{T}_i\}_{i=1}^K$  corresponding to the  $K$  attention mask maps.  $\mathbf{T}_i \in \mathbb{R}^{C \times H \times Q}$  can be computed by the following equation:

$$\mathbf{T}_i = R(\mathbf{M}_i) \odot \mathbf{X}_i^{\text{CNN}} \quad (7)$$

where  $R(\cdot)$  represents the expansion of  $\mathbf{M}_i$  into the same size as  $\mathbf{X}_i^{\text{CNN}}$  and  $\odot$  represents element-by-element multiplication.

(2) *Visual-semantic embedding encoder*: The visual-semantic embedding encoder is used to encode  $K$  attention feature maps into local visual embeddings. This encoder consists primarily of a convolution-pooling structure and a linear transformation layer. First, the attention feature maps  $\mathbf{T}_i$  ( $i = 1, \dots, K$ ) are concatenated along the channel dimension to obtain the concatenated attention feature map  $\mathbf{\Gamma} \in \mathbb{R}^{KC \times H \times Q}$ :

$$\mathbf{\Gamma} = \text{Concat}(\mathbf{T}_i) \quad (8)$$

where  $\text{Concat}(\cdot)$  represents the concatenation operation.

Then, the  $C \times 1$  convolutions are used to perform convolution operations with  $\Gamma$ , and global average pooling is performed to obtain:

$$\mathbf{P} = \text{Avgpooling}(\text{Conv}_{1 \times 1}(\Gamma)) \quad (9)$$

Finally,  $\mathbf{P}$  is multiplied by a learnable linear transform factor  $\mathbf{W} \in \mathbb{R}^{C \times K}$  to obtain the local visual embeddings, which can be expressed as follows:

$$\Phi_{\text{LVFE}}(\mathbf{X}_i) = \mathbf{P} \times \mathbf{W} \quad (10)$$

Similarly, local visual embeddings are obtained using a loss function based on cross-entropy and guided by semantic knowledge. This function is defined as follows:

$$L_{\text{LVFE}} = -\frac{1}{N_b} \sum_{j=1}^{N_b} y_j \log \frac{\exp(\Phi_{\text{LVFE}}(\mathbf{X}_i) \times \varphi(y_j))}{\sum_{l \in \mathcal{Y}} \exp(\Phi_{\text{LVFE}}(\mathbf{X}_i) \times \varphi(y_l))} \quad (11)$$

### 3.3.4. Joint Optimization of the GLVFENet and Prediction

The full model is jointly trained in an end-to-end manner. Therefore, the shared feature extraction module, GVFE-Subnet, and LVFE-Subnet are simultaneously optimized using the following objective function:

$$L = L_{\text{GVFE}} + \lambda \cdot L_{\text{LVFE}} \quad (12)$$

where  $\lambda$  is a hyperparameter. The optimal parameters of the network can be obtained by minimizing the loss function.

After the model is trained, we fuse global and local embeddings and use the fusion results to predict the test samples. First, we obtain the embeddings of the test samples in the semantic spaces of GVFE-Subnet and LVFE-Subnet, denoted as  $\Phi_{\text{GVFE}}(\mathbf{X}_i)$  and  $\Phi_{\text{LVFE}}(\mathbf{X}_i)$ , respectively. Then, we use a pair of fusion coefficients ( $\alpha_1$  and  $\alpha_2$ ) to combine these embeddings. Finally, the label of the input sample is predicted by matching it with the semantic vectors of each class and applying a calibrated stacking method, as shown in the following equation:

$$\hat{y} = \underset{\tilde{y} \in \mathcal{Y}}{\text{argmax}} (\alpha_1 \Phi_{\text{GVFE}}(\mathbf{X}_i) + \alpha_2 \Phi_{\text{LVFE}}(\mathbf{X}_i)) \varphi(\tilde{y}) - \gamma \mathbb{I}[\tilde{y} \in \mathcal{Y}^s] \quad (13)$$

where  $\mathbb{I}[\tilde{y} \in \mathcal{Y}^s]$  is the indicator function (i.e.,  $\mathbb{I} = 1$  if  $\tilde{y}$  is a known class and 0 otherwise), and  $\gamma$  is a calibration factor adjusted on the validation set.

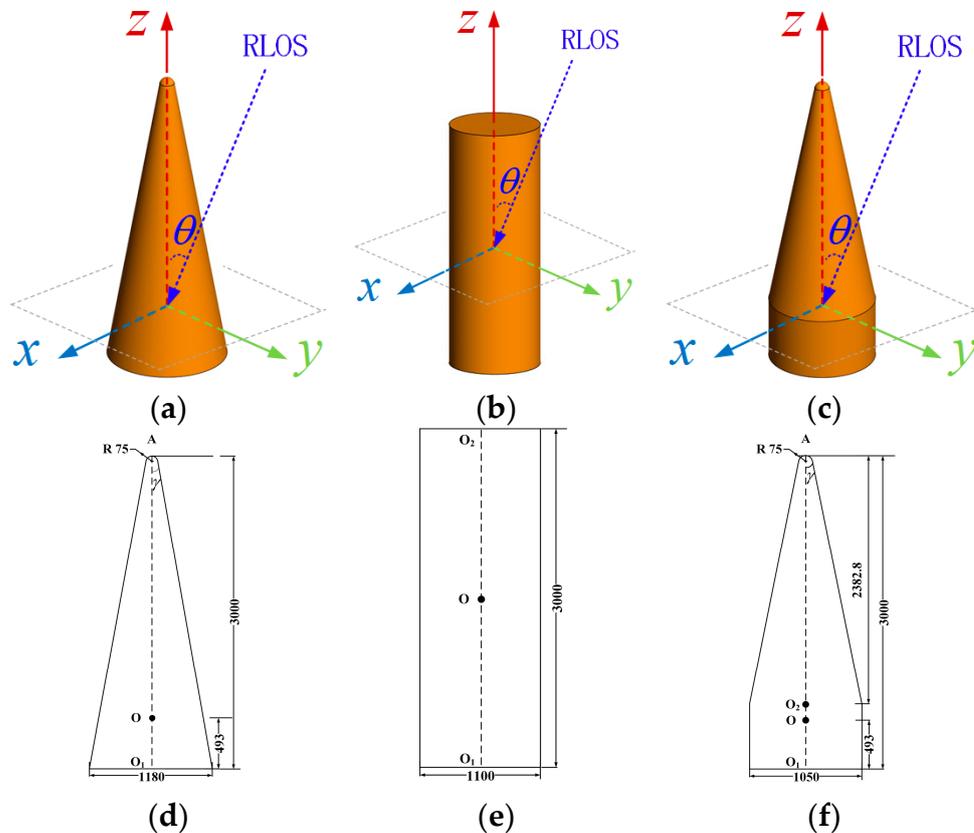
## 4. Results

### 4.1. Data Generation

Due to the difficulty of obtaining measured data on real space targets [2,5,52], most of the relevant research in this field use simulation data. The generation method of dynamic radar echoes based on EM calculation data is as follows: Static EM calculation data are first generated for space targets at all attitude angles via the FEKO EM calculation software. Then, based on the target's micromotion model, we obtain the attitude angle sequence of the target in real-time, which is relative to the radar line-of-sight (RLOS). Finally, the dynamic radar echo of the target can be obtained by querying static EM calculation data with real-time attitude angle sequences. To obtain the HRRP sequence, we adopted the Fast Fourier transform (FFT) in MATLAB 2017a to process the dynamic echoes.

We selected three typical space targets. Figure 2 depicts their three-dimensional CAD models and sectional drawings. The surfaces of the targets are coated with the same material. The micromotion forms of these targets include precession, nutation, wobble, and tumble. Usually, only tumble is thought to be possible with cylinder-shaped targets. Therefore, we generate nine categories of targets by combining the three geometric shapes

and four micromotion forms while considering that the cylinder-shaped target just tumbles. Table 2 provides a detailed description of these categories, and Table 3 lists the four types of micromotion parameters.



**Figure 2.** CAD models and sectional drawings of cone-shaped target, cylinder-shaped target, and cone-cylinder-shaped target. (a–c) are their CAD models. (d–f) are their sectional drawings.

**Table 2.** Training validation and test set details.

Target Type	Target Shape	Micromotion Type	Training Set Size	Validation Set Size	Testing Set Size
T1	Cone	Precession	800	100	100
T2	Cone	Nutation	800	100	100
T3	Cone	Wobble	800	100	100
T4	Cone	Tumble	800	100	100
T5	Cylinder	Tumble	800	100	100
T6	Cone-cylinder	Precession	800	100	100
T7	Cone-cylinder	Nutation	800	100	100
T8	Cone-cylinder	Wobble	800	100	100
T9	Cone-cylinder	Tumble	800	100	100
Sum	/	/	7200	900	900

**Table 3.** Micromotion parameter settings.

m-D Parameter	Precession	Nutation	Wobble	Tumble
$f_s$ (Hz)	5	3	/	/
$f_c$ (Hz)	2:0.1:3.9	2:0.5:3.5	/	/
$f_w$ (Hz)	/	1:0.25:2	1:0.25:2	0.5:0.025:0.975
$A_w$ (°)	/	5	3:2:9	1

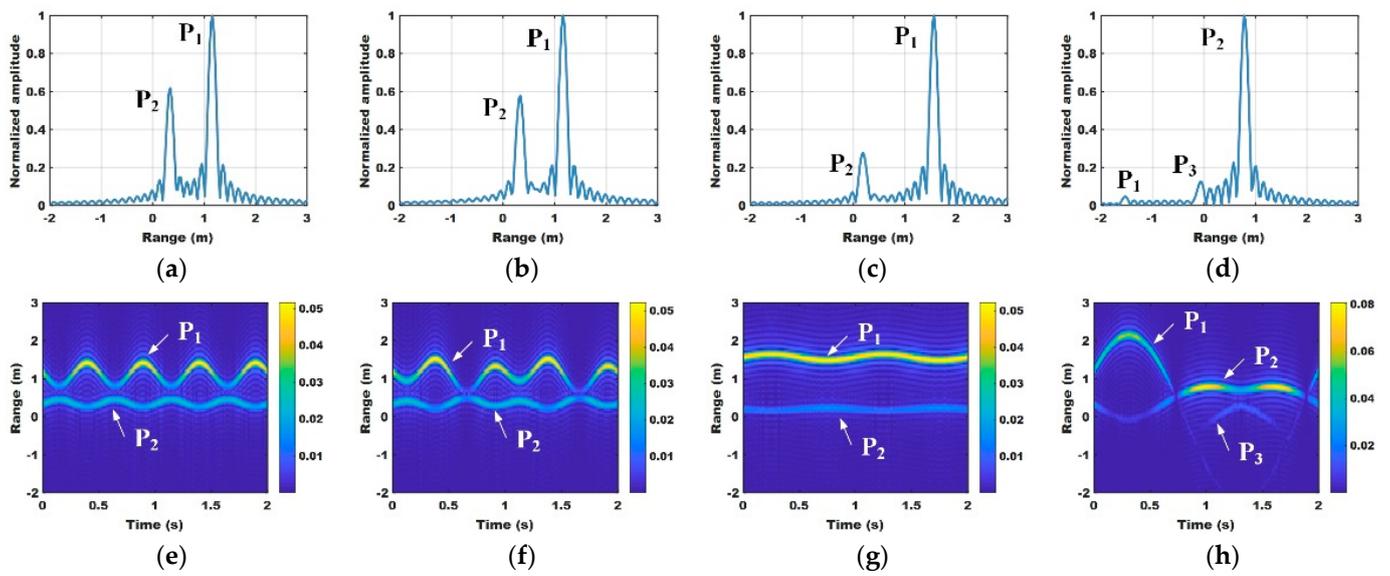
The radar center frequency is set to 10 GHz, with a working bandwidth of 1 GHz and a frequency step of 10 MHz, using a polarization mode of horizontal transmission and horizontal reception. The physical optics method is adopted to calculate the static EM calculation data of the space targets. Since the three geometries used in this article are rotational symmetric targets, we set the angle between the plane wave's incident direction and the target's longitudinal symmetry axis to change from  $0^\circ$  to  $360^\circ$  at  $0.01^\circ$  intervals. In this way, static EM calculation data of space targets at all attitude angles can be obtained in FEKO. To obtain the dynamic HRRP sequence of space targets, the pulse repetition frequency (PRF) is set to 1000 Hz, the pulse dwell time is 2 s, the elevation angle of the RLOS is set to change from  $25^\circ$  to  $75^\circ$  at  $1^\circ$  intervals, the direction of the target rotation axis is (0, 0, 1), and the initial Euler angle is (20, 3, 45). These parameters are summarized in Table 4.

**Table 4.** RLOS direction and parameter settings.

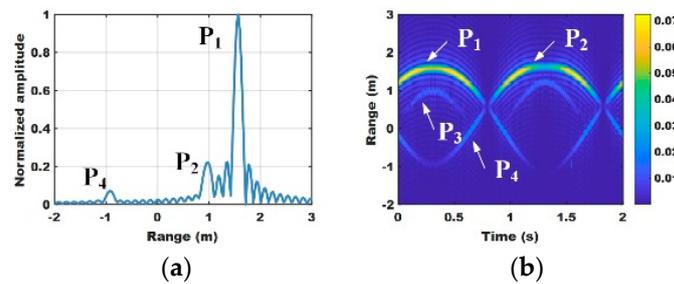
Parameter	Value	Step
Carrier frequency (GHz)	10	/
Bandwidth (GHz)	1	0.01
Pulse repetition frequency (Hz)	1000	/
Observation time (s)	2	/
RLOS elevation (°)	[25,75]	$1^\circ$
Coning axis	(0, 0, 1)	/
Initial Euler angle	(20, 3, 45)	/

To simulate radar data in real scenes, we added noise with an SNR = 5 dB to the radar echoes. Based on the above target structure, motion, and radar parameters, we generate dynamic HRRP sequences for nine categories of targets. Figures 3–5 depict one sample from each target category. The dimension of a single HRRP sequence is  $512 \times 2000$  (FFT of 512 sampling points is adopted), and the sample size for each category is 1000. In the experiment, we randomly divided the samples into the training set, validation set, and testing set, and their ratio is 8:1:1. More specifically, for seen classes, the sample sizes of the training set, validation set, and testing set account for 80%, 10%, and 10% of the total number of samples in each category, respectively. For unseen classes, the sample sizes of the validation and testing sets accounted for 10% and 10% of the sample size of each category, respectively.

To validate the performance of the proposed GLVFENet, we design two division methods for seen and unseen classes. Detailed information about these methods listed in Table 5. The division method is based mainly on the following two principles: (1) To ensure that semantics can be transferred from seen classes to unseen classes, the seen class samples include the attributes of unseen class samples. (2) In the same division, the number of unseen classes shows an increasing trend, and the categories of these unseen class samples exhibit a hierarchical relationship. In different divisions, the unseen class samples exhibit two different shapes.



**Figure 3.** HRRP and HRRP sequence of a cone-shaped target with different micromotion forms. The first to fourth columns represent precession, nutation, wobble, and tumble, respectively. The first row represents the HRRP of the cone-shaped target with different micro-motion forms. The second row represents the HRRP sequences of the cone-shaped target with different micro-motion forms.



**Figure 4.** HRRP and HRRP sequence of a cylinder-shaped target with tumble micromotion form. (a) is the HRRP of cylinder-shaped target. (b) is the HRRP sequence of the cylinder-shaped target.

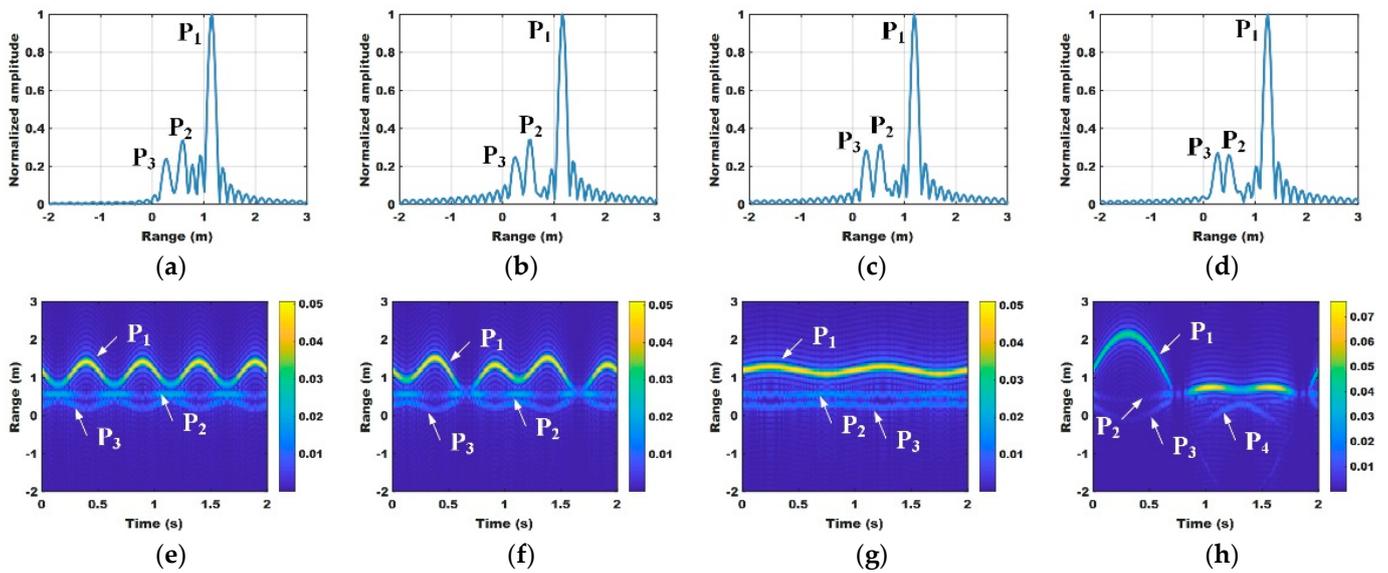
**Table 5.** Division of seen and unseen classes.

$N_s/N_u$	Target Type	Division I	Division II
8/1	T1	Seen	Seen
	T2	Seen	Seen
	T3	Unseen	Seen
	T4	Seen	Seen
	T5	Seen	Seen
	T6	Seen	Seen
	T7	Seen	Seen
	T8	Seen	Unseen
	T9	Seen	Seen
7/2	T1	Seen	Seen
	T2	Seen	Seen
	T3	Unseen	Seen
	T4	Unseen	Seen
	T5	Seen	Seen
	T6	Seen	Seen
	T7	Seen	Seen
	T8	Seen	Unseen
	T9	Seen	Unseen

Table 5. Cont.

$N_s/N_u$	Target Type	Division I	Division II
6/3	T1	Seen	Seen
	T2	Seen	Unseen
	T3	Unseen	Seen
	T4	Unseen	Seen
	T5	Seen	Seen
	T6	Seen	Seen
	T7	Unseen	Seen
	T8	Seen	Unseen
	T9	Seen	Unseen

$N_s$  and  $N_u$  represent the number of seen and unseen classes, respectively.



**Figure 5.** HRRP and HRRP sequence of a cone–cylinder-shaped target with different micromotion forms. The first to fourth columns represent precession, nutation, wobble, and tumble, respectively. The first row represents the HRRP of the cone-cylinder-shaped target with different micro-motion forms. The second row represents the HRRP sequences of the cone-cylinder-shaped target with different micro-motion forms.

#### 4.2. Evaluation Metrics

Following the commonly used evaluation metrics in GZSL, we adopt the average top-1 accuracy of each class as the evaluation criterion, represented by the following formula:

$$\text{Acc} = \frac{1}{\|Y\|} \sum_{T=1}^P \frac{P}{Q} \quad (14)$$

where  $P$  represents the number of correctly recognized samples for a certain class,  $Q$  represents the total number of samples for that class,  $Y$  represents the class labels, and  $\|Y\|$  represents the number of classes. We denote top-1 accuracies of the seen and unseen classes as  $\text{Acc}_s$  and  $\text{Acc}_u$ , respectively. To comprehensively reflect the recognition performance of the network for seen and unseen classes, we use the harmonic mean as an indicator. The harmonic mean is represented by the following formula:

$$H = \frac{2 \times \text{Acc}_s \times \text{Acc}_u}{\text{Acc}_s + \text{Acc}_u} \quad (15)$$

To further measure the performance of the model, we also introduce three metrics that are model parameters (Params), model running time (Time, in seconds), and floating-point operations per second (FLOPs).

#### 4.3. Implementation Details

In the shared feature extraction module, the CNN backbone network adopts ResNet101, which improves HRRP sequence representation through fine-tuning in an end-to-end way. The initial learning rate is set to  $1 \times 10^{-4}$ , and after every 10 epochs, the learning rate decays by 0.8 times its original value. A total of 100 epochs were conducted, with a batch size of 8 for each epoch. The AdamW optimizer is used to minimize the loss function. All the experiments are performed on the Ubuntu 18.06 operating system, utilizing an Intel Xeon CPU E5-2698v4 equipped with an NVIDIA Tesla V100 32-GB GPU. The development environment is Python 3.9 with PyTorch 1.7.0.

#### 4.4. Experimental Results

Due to the lack of ZSL methods for space target recognition in this field, we compare several popular methods in the computer vision field. We conducted extensive comparative experiments on the two division methods mentioned in Table 5. The comparison results are shown in Tables 6 and 7. These results exhibit the following: (1) As the number of unseen classes increases in the testing set, the  $H$  for all methods decreases. This is reasonable because as the number of unseen classes increases, recognition becomes more challenging. (2) When the number of unseen classes is the same, the proposed method, GLVFENet, consistently outperforms the other methods in terms of  $H$ , regardless of the division method. Table 6 demonstrates that our proposed GLVFENet outperforms the best-compared method APN in terms of  $H$ , improving by 8.8%, 12.4%, and 18.1% when the number of unseen classes is 1, 2, and 3, respectively. Table 7 reveals that our proposed GLVFENet outperforms the best-compared method APN in terms of  $H$ , improving by 15.6%, 13.1%, and 10.1% when the number of unseen classes is 1, 2, and 3, respectively. (3) Further analysis reveals that when the number of unseen classes is the same, our method performs better in Division I (Table 6) and achieves a higher  $H$  value than Division II (Table 7), with improvements of 1.0%, 1.3%, and 5.8%, respectively. This is understandable because the unseen classes of Division II have more complex geometric shapes and EM scattering characteristics than those of Division I, increasing the difficulty of target recognition. In conclusion, our method achieves the best results in generalized zero-shot space target recognition, demonstrating that our approach effectively aligns visual features with semantic knowledge by utilizing cooperative global and local visual features simultaneously. Our approach enhances the transferability of semantic knowledge from seen classes to unseen classes, thereby improving zero-shot recognition performance.

We have made statistics on Params, FLOPs, and Time. The statistical results are shown in Table 8. From Table 8, although the proposed GLVFENet has the best recognition accuracy, its parameters and FLOPs are relatively high. This might be because the LVFE-Subnet utilizes convolutional operations to encode the  $K$  attention feature into local visual embeddings, which involves a larger number of input and output channels. As a result, the overall network has higher parameters. However, since convolutional operations are performed in parallel for each channel, the inference time has not significantly increased. In the future, our focus will be on researching lightweight network structures that can achieve a high recognition accuracy while using fewer network parameters.

**Table 6.** Recognition accuracy comparison among different networks (Division I).

Method	8/1			7/2			6/3		
	Acc <sub>s</sub>	Acc <sub>u</sub>	H	Acc <sub>s</sub>	Acc <sub>u</sub>	H	Acc <sub>s</sub>	Acc <sub>u</sub>	H
HSVA [18]	71.90	95.40	82.00	44.70	95.90	61.00	33.1	98.60	49.50
APN [22]	98.20	83.90	90.50	81.20	80.00	80.60	60.20	58.30	59.20
DPPN [23]	62.19	61.66	61.93	50.67	49.80	50.23	36.27	33.70	34.94
GLVFENet	99.70	98.90	99.30	96.40	89.70	93.00	93.50	75.90	77.30

**Table 7.** Recognition accuracy comparison among different networks (Division II).

Method	8/1			7/2			6/3		
	Acc <sub>s</sub>	Acc <sub>u</sub>	H	Acc <sub>s</sub>	Acc <sub>u</sub>	H	Acc <sub>s</sub>	Acc <sub>u</sub>	H
HSVA [18]	61.80	97.50	75.70	61.10	50.00	55.00	39.30	65.40	49.10
APN [22]	94.60	73.40	82.70	81.50	76.00	78.60	81.20	49.30	61.40
DPPN [23]	52.80	49.76	51.23	69.92	33.75	45.53	39.71	37.40	38.52
GLVFENet	99.60	97.10	98.30	98.10	86.10	91.70	80.80	64.10	71.50

**Table 8.** Performance display of different method.

Method	Params	FLOPs	Time (s)
HSVA [18]	4,458,624	8,916,992	0.004059
APN [22]	42,500,160	159,948,603,392	0.016310
DPPN [23]	45,176,699	8,851,663,192	0.021384
GLVFENet	101,252,617	219,169,602,770	0.017393

## 5. Discussion

In this section, we take 7/2 in Table 5 (Division I) as an example for further discussion, including visualization, an ablation study, and hyperparameter analysis.

### 5.1. Visualization

#### 5.1.1. Confusion Matrix

To intuitively illustrate the classification performance of different methods, we present their confusion matrices as shown in Figure 6. In the confusion matrix, each row represents the true class label of the target, and each column represents the class label predicted by the model. The squares on the main diagonal represent correctly classified samples. A darker color indicates that more samples are correctly classified into this category than that represented by a lighter color. Figure 6 shows that GLVFENet has the best overall recognition performance and performs well in recognizing various categories of targets.

#### 5.1.2. Feature Cluster Analysis

To provide more evidence of the efficacy of our GLVFENet, we employ the t-SNE distribution [53] to analyze the feature clustering of nine categories of space targets. This analysis is visually shown in Figure 7. In Figure 7, each point represents a sample, and each color represents a class. Our method exhibits better intraclass compactness and interclass separability than the other methods. The GLVFENet, through learning both global and local features, enables the model to capture more comprehensive and discriminative features, thereby improving the model's recognition capability.

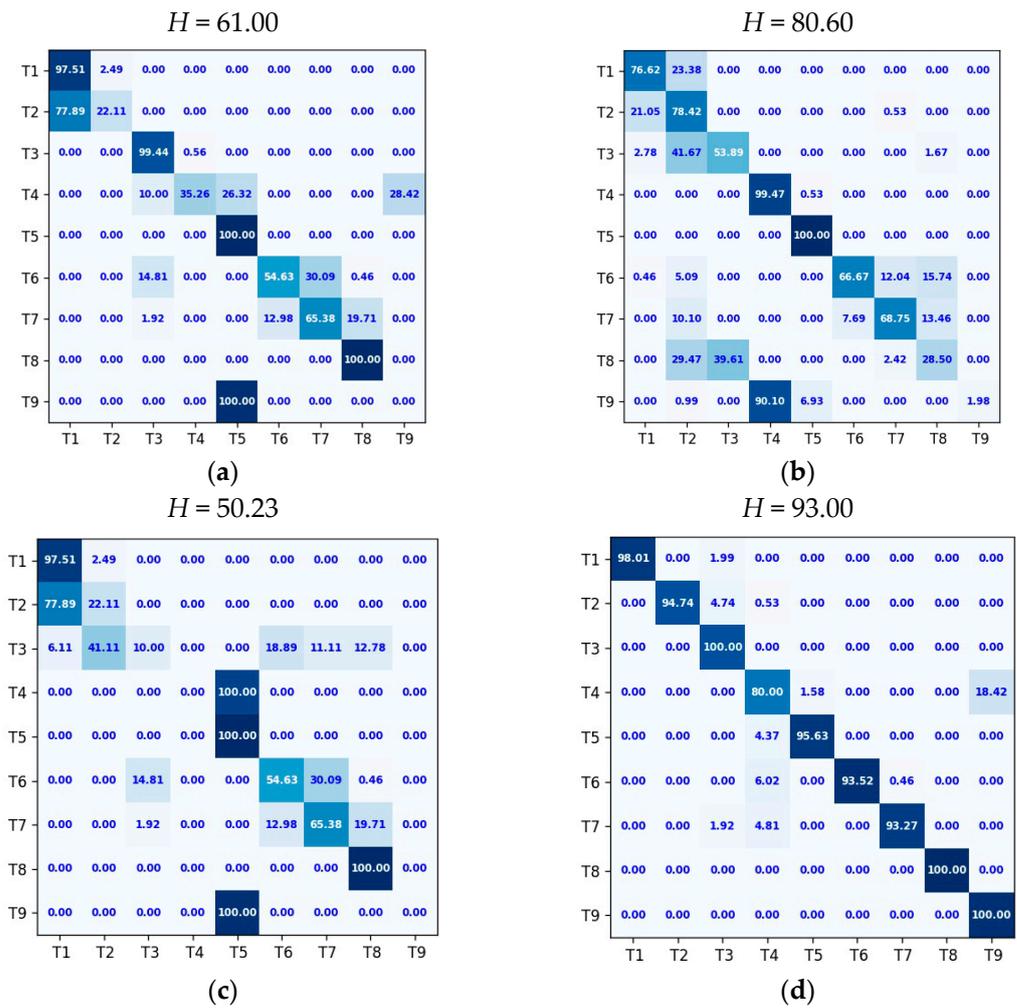


Figure 6. Confusion matrices for different methods. (a) HSVA, (b) APN, (c) DPPN, (d) GLVFENet.

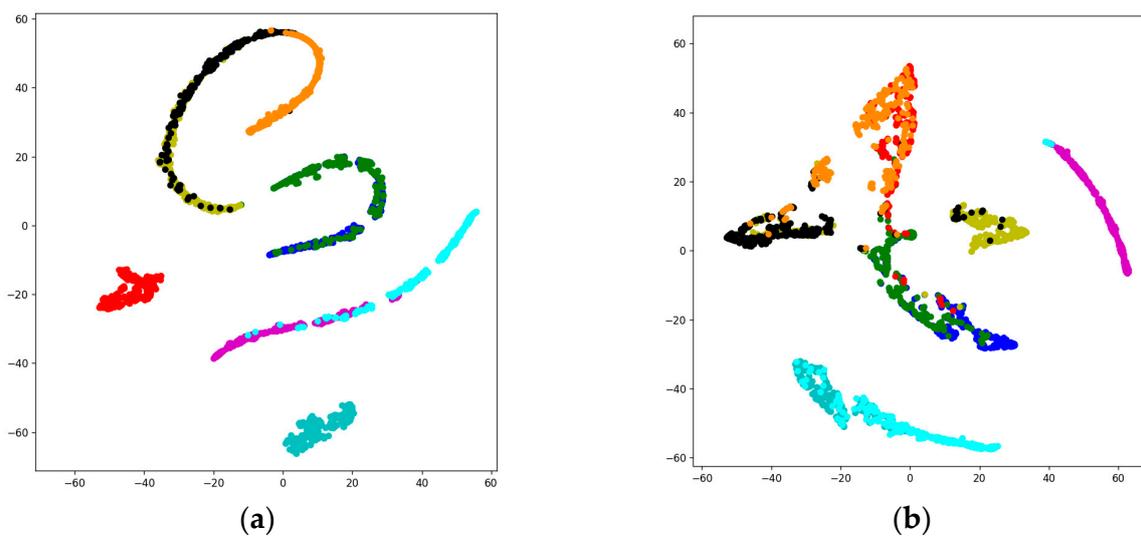


Figure 7. Cont.

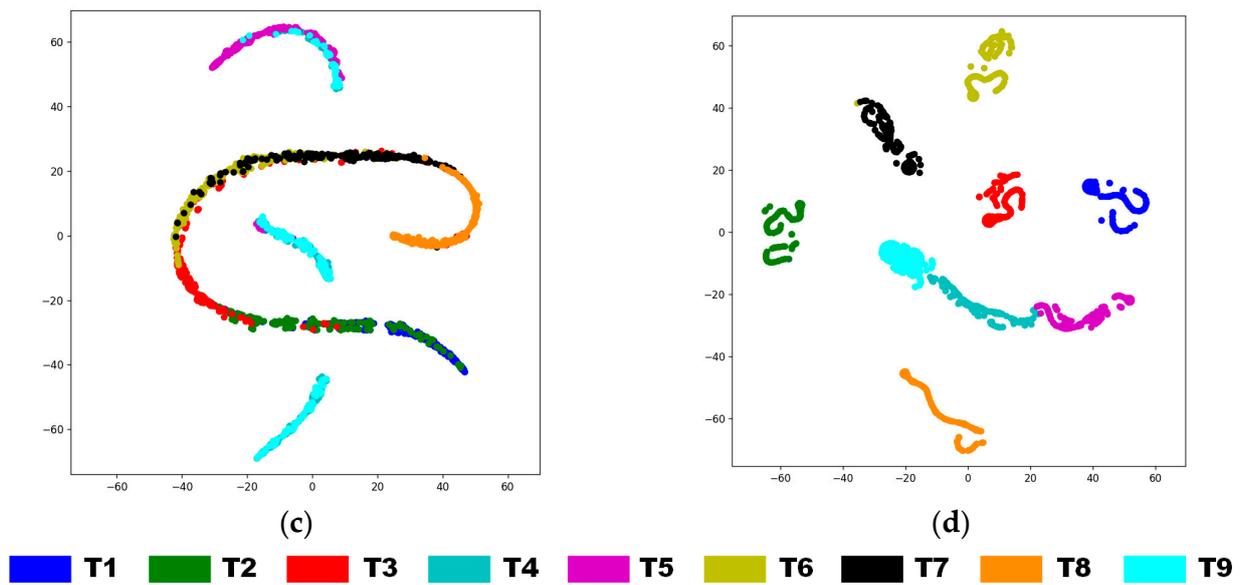


Figure 7. 2-D t-SNE visualization of different methods. (a) HSVA, (b) APN, (c) DPPN, (d) GLVFENet.

### 5.1.3. ROC Curves

Recognition accuracy can reflect the relationship between the number of correctly classified test samples and the total number of test samples. However, it does not reflect the relationship between misclassification rate (the probability of classifying a false sample as true) and sensitivity (the probability of classifying a true sample as true). Therefore, we evaluate the recognition effect by using ROC curves and AUC. The closer the ROC curve is to the upper left corner of the graph, the better the classification performance of one model, and the corresponding AUC value of one model is close to 1. Figure 8 shows the ROC curves and corresponding AUC values of the comparison methods under the 7/2 condition in Table 5. The ROC curve of GLVFENet is closer to the upper left corner than the ROC curves of the other models. The corresponding AUC value is also larger, indicating that GLVFENet performs well in zero-shot classification tasks.

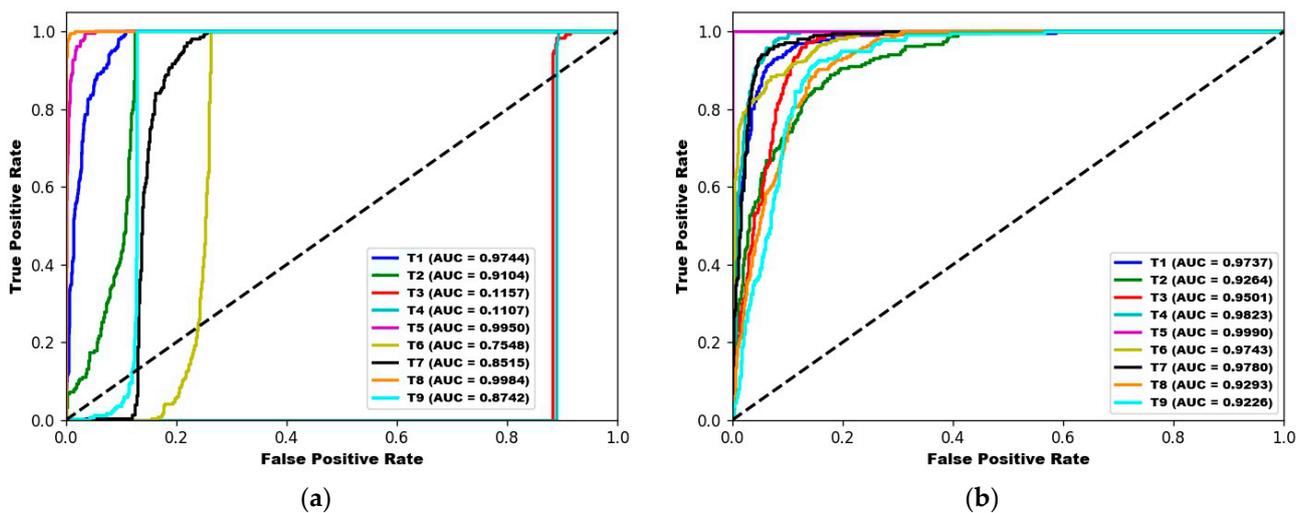
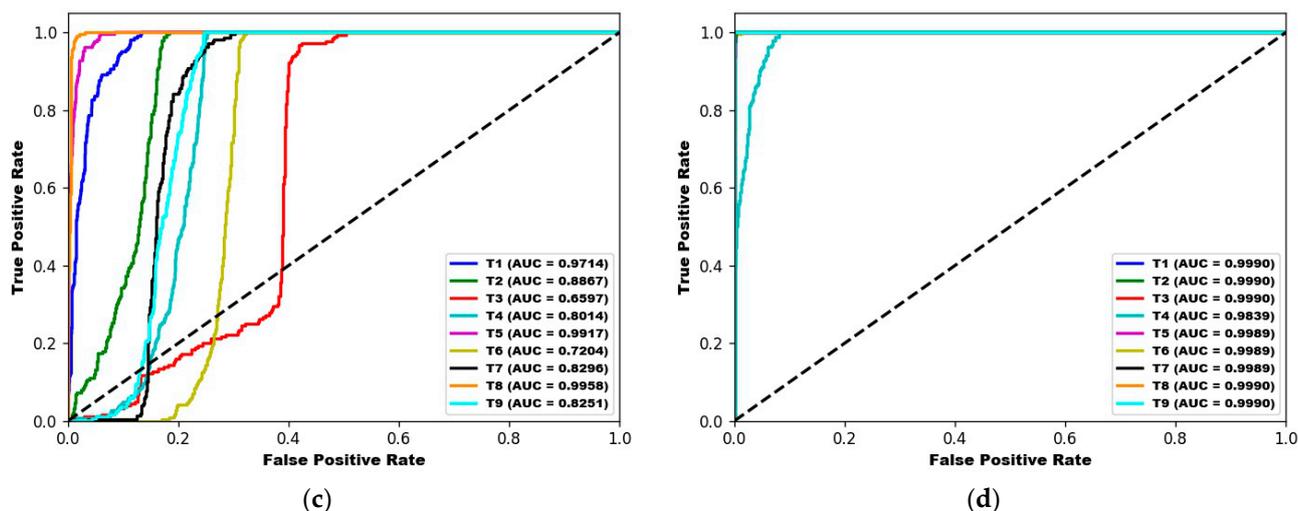


Figure 8. Cont.



**Figure 8.** ROC curves and AUC of different methods. (a) HSVA, (b) APN, (c) DPPN, (d) GLVFENet.

### 5.2. Ablation Study

To validate the impact of each component on the model, we conduct ablation experiments. Table 9 displays the results of these experiments. “GLVFENet w/o LVFE-Subnet” indicates that only GVFE-Subnet is used. “GLVFENet w/o GVFE-Subnet” indicates that only the LVFE-Subnet is used. “GLVFENet w/o cosine similarity” indicates the cosine similarity is removed from the GVFE-Subnet loss function and only the cross-entropy loss is utilized. “GLVFENet w/o  $L_{CS}$ ” indicates the calibrated stacking method is not used during prediction. Table 8 illustrates that (1) GLVFENet outperforms GVFE-Subnet and LVFE-Subnet by 79.6% and 27.3% in terms of  $H$ , respectively. This demonstrates the importance of both global and local features in GZSL recognition. In particular, the semantically enhanced local features have a greater influence on recognition accuracy. (2) Using the cosine similarity in the classification loss function to calculate the compatibility score between the global visual projection and semantic vectors results in an 18.1% improvement. This is because cosine similarity can constrain and reduce the variance of neurons, improving the generalizability of the model. (3) The calibrated stacking constraint helps alleviate the tendency of unseen classes being recognized as seen classes during GZSL, thus improving the recognition accuracy of unseen classes and overall recognition accuracy. Therefore, each component in the model improves the overall performance of the model.

**Table 9.** Ablation Studies for Different Components of GLVFENet (Division I).

Methods	$Acc_s$	$Acc_u$	$H$
GLVFENet w/o LVFE-Subnet	12.0	15.0	13.4
GLVFENet w/o GVFE-Subnet	99.7	48.6	65.7
GLVFENet w/o cosine similarity	63.0	92.3	74.9
GLVFENet w/o $L_{CS}$	99.2	54.1	70.0
GLVFENet (full)	96.4	89.7	93.0

### 5.3. Hyperparameter Analysis

#### 5.3.1. Effect of Scaling Factor $\sigma$

$\sigma$  denotes the scaling factor in the loss of GVFE-Subnet, and Figure 9 illustrates the effect of varying  $\sigma$  over the range  $\{2, 5, 10, 15, 20, 25\}$  on the model performance. As  $\sigma$  gradually increases,  $Acc_s$ ,  $Acc_u$ , and  $H$  increase until  $\sigma = 10$ , where  $Acc_s$ ,  $Acc_u$ , and  $H$  reach their highest values of 96.4%, 89.7%, and 93.0%, respectively. As  $\sigma$  further increases,  $Acc_s$ ,  $Acc_u$ , and  $H$  show a downward trend. For this reason, we choose  $\sigma = 10$ , when the model performs best.

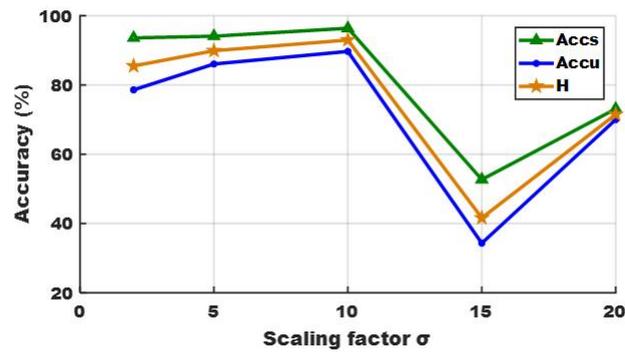


Figure 9. Effect of scaling factor  $\sigma$  in  $L_{GVFE}$ .

### 5.3.2. Effect of Number of $K$ in LVFE-Subnet

$K$  denotes the number of mask attention maps generated by LVFE-Subnet, which is an important hyperparameter. We take values for  $K$  within the range  $\{4, 6, 8, 10, 12, 14\}$ . The experimental results are shown in Figure 10. When  $K = 8$ , GLVFENet performs best, achieving the highest accuracy for seen and unseen classes. Therefore, we set  $K = 8$ .

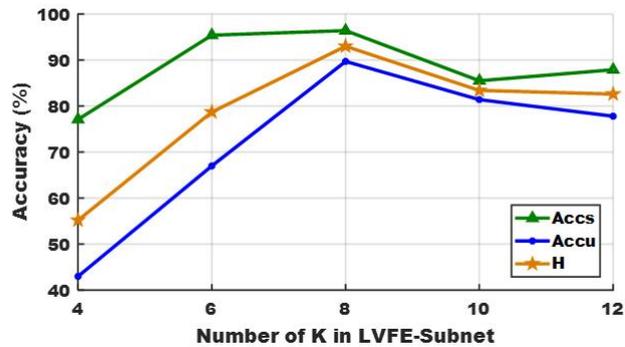


Figure 10. Effect of the number of attention mask maps.

### 5.3.3. Effect of Weight $\lambda$ in Loss Function

We vary  $\lambda$  in (12) within the range of  $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ , as shown in Figure 11. As  $\lambda$  increases,  $H$  gradually increases until  $\lambda = 0.8$ , when our method performs best. However, as  $A$  continues to increase,  $H$  decreases. Therefore, we choose  $\lambda = 0.8$ .

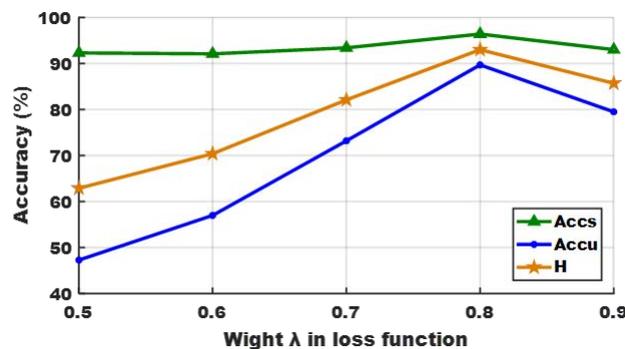


Figure 11. Effect of the weight ( $\lambda$ ) in the loss function.

### 5.3.4. Effect of Calibration Factor $\gamma$

To alleviate the tendency of unseen classes to be classified as seen classes, we introduce the calibrated stacking method into GLVFENet during inference. We control the calibrated stacking constraint in (13) via  $\gamma$ . As shown in Figure 12, as  $\gamma$  increases, the accuracy of seen classes starts to decrease, while the accuracy of unseen classes continues to increase. This indicates that the calibrated stacking constraint is effective in our model. When  $\gamma = 0.999$ ,

$H$  achieves the best result of 93.0%. As  $\gamma$  continues increasing,  $H$  decreases. Therefore, we choose  $\gamma = 0.999$ .

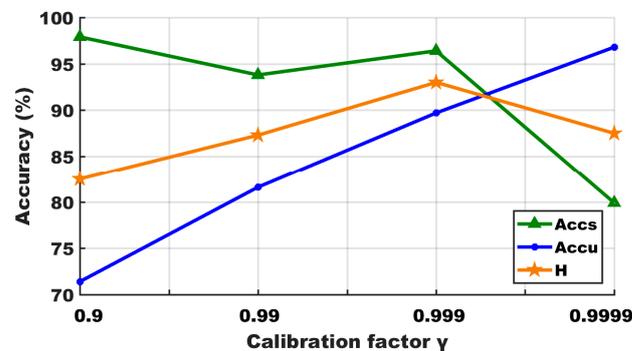


Figure 12. Effect of the calibration factor  $\gamma$ .

### 5.3.5. Effect of Combination Coefficients ( $\alpha_1, \alpha_2$ )

To determine the effect of combining factors ( $\alpha_1, \alpha_2$ ) on the model, we varied ( $\alpha_1, \alpha_2$ ) from  $\{(0.1, 0.9), (0.2, 0.8), (0.3, 0.7), (0.4, 0.6), (0.5, 0.5), (0.6, 0.4), (0.7, 0.3), (0.8, 0.2), (0.9, 0.1)\}$ . As shown in Figure 13, when the values of  $\alpha_1/\alpha_2$  are overly large or small, GLVFENet performs poorly. When ( $\alpha_1, \alpha_2$ ) = (0.6, 0.4), GLVFENet achieves the best performance, indicating that the model fully leverages the important roles of both global and local features.

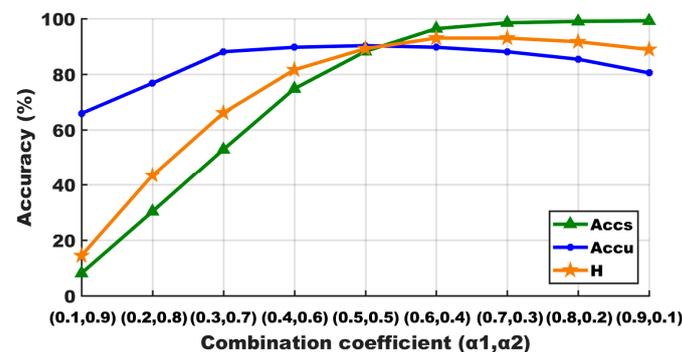


Figure 13. Effect of the combination coefficient ( $\alpha_1, \alpha_2$ ) between GVFE-Subnet and LVFE-Subnet.

## 6. Conclusions

In this paper, we propose a space target recognition network based on GZSL, named GLVFENet. GLVFENet consists primarily of GVFE-Subnet and LVFE-Subnet. The GVFE-Subnet based on cosine metric learning is mainly used to obtain coarse-grained global embedding; the LVFE-Subnet based on soft space attention is mainly used to capture local embedding in images. This dual-branch network jointly learns global and local visual features to obtain discriminative target features. During the testing phase, the calibrated stacking method is introduced to alleviate the tendency toward misclassifying unseen classes as seen classes. In addition, we incorporate domain-prior knowledge to design binary semantic attributes of space targets. Extensive experiments validate the effectiveness of our proposed method.

Future work will focus on further improving the accuracy of recognizing unseen classes of space targets, while optimizing the network to reduce parameters and complexity.

**Author Contributions:** Conceptualization, methodology, and software, Y.Z. and J.G.; validation, H.W. and K.L.; investigation, K.L. and Q.Z.; data curation, Y.L.; writing—original draft preparation, Y.Z. and J.G.; writing—review and editing, Y.Z., J.G., K.L., Y.L. and Q.Z.; supervision, K.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Nature Science Foundation of China under Grant 62371468, 62131020 and 62301599.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank all reviewers and editors for their comments on this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Tian, X.D.; Bai, X.R.; Zhou, F. Recognition of Micro-Motion Space Targets Based on Attention-Augmented Cross-Modal Feature Fusion Recognition Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5104909. [[CrossRef](#)]
2. Zhang, R.Z.; Wang, Y.; Yeh, C.M.; Lu, X.F. Precession Parameter Estimation of Warhead with Fins Based on Micro-Doppler Effect and Radar Network. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 443–459. [[CrossRef](#)]
3. Choi, O.; Park, S.H.; Kim, M.; Kang, K.B.; Kim, K.T. Efficient Discrimination of Ballistic Targets with Micromotions. *IEEE Trans. Aerosp. Electron. Syst.* **2020**, *56*, 1243–1261. [[CrossRef](#)]
4. Persico, A.R.; Clemente, C.; Gaglione, D.; Ilioudis, C.V.; Cao, J.; Pallotta, L.; De Maio, A.; Proudler, I.K.; Soraghan, J.J. On Model, Algorithms, and Experiment for Micro-Doppler-Based Recognition of Ballistic Targets. *IEEE Trans. Aerosp. Electron. Syst.* **2017**, *53*, 1088–1108. [[CrossRef](#)]
5. Chen, X.B.; Ye, C.M.; Wang, Y.; Zhang, Y.; Hu, Q.R. Unambiguous Estimation of Multidimensional Parameters for Space Precession Targets with Wideband Radar Measurements. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5112716. [[CrossRef](#)]
6. Liu, L.H.; McLernon, D.; Ghogho, M.; Hu, W.D.; Huang, J. Ballistic missile detection via micro-Doppler frequency estimation from radar return. *Digit. Signal Process.* **2012**, *22*, 87–95. [[CrossRef](#)]
7. Chen, J.; Xu, S.Y.; Chen, Z.P. Convolutional neural network for classifying space target of the same shape by using RCS time series. *IET Radar Sonar Navig.* **2018**, *12*, 1268–1275. [[CrossRef](#)]
8. Wang, Y.Z.; Feng, C.Q.; Hu, X.W.; Zhang, Y.S. Classification of Space Micromotion Targets with Similar Shapes at Low SNR. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3504305. [[CrossRef](#)]
9. Ye, L.; Hu, S.B.; Yan, T.T.; Meng, X.; Zhu, M.Q.; Xu, R.Z. Radar target shape recognition using a gated recurrent unit based on RCS time series' statistical features by sliding window segmentation. *IET Radar Sonar Navig.* **2021**, *15*, 1715–1726. [[CrossRef](#)]
10. Wang, L.; Bai, X.R.; Gong, C.; Zhou, F. Hybrid Inference Network for Few-Shot SAR Automatic Target Recognition. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9257–9269. [[CrossRef](#)]
11. Pourpanah, F.; Abdar, M.; Luo, Y.X.; Zhou, X.L.; Wang, R.; Lim, C.P.; Wang, X.Z.; Wu, Q.M.J. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4051–4070. [[CrossRef](#)] [[PubMed](#)]
12. Chen, S.M.; Hong, Z.M.; Liu, Y.; Xie, G.S.; Sun, B.G.; Li, H.; Peng, Q.M.; Lu, K.; You, X.G. TransZero: Attribute-Guided Transformer for Zero-Shot Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 330–338.
13. Romera-Paredes, B.; Torr, P. An embarrassingly simple approach to zero-shot learning. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 2152–2161.
14. Shigetou, Y.; Suzuki, I.; Hara, K.; Shimbo, M.; Matsumoto, Y. Ridge regression, hubness, and zero-shot learning. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, Porto, Portugal, 7–11 September 2015; pp. 135–151.
15. Zhang, L.; Wang, P.; Liu, L.Q.; Shen, C.H.; Wei, W.; Zhang, Y.N.; van den Hengel, A. Towards Effective Deep Embedding for Zero-Shot Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 2843–2852. [[CrossRef](#)]
16. Xian, Y.Q.; Lorenz, T.; Schiele, B.; Akata, Z. Feature generating networks for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5542–5551.
17. Han, Z.Y.; Fu, Z.Y.; Chen, S.; Yang, J. Contrastive Embedding for Generalized Zero-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–21 June 2021; pp. 2371–2381.
18. Chen, S.M.; Xie, G.S.; Liu, Y.; Peng, Q.M.; Sun, B.G.; Li, H.; You, X.G.; Shao, L. HSVA Hierarchical Semantic-Visual Adaptation for Zero-Shot Learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16622–16634.
19. Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; Akata, Z. Generalized zero- and few-shot learning via aligned variational Autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8239–8247.
20. Xie, G.S.; Liu, L.; Jin, X.B.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.Z.; Shao, L. Attentive region embedding network for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 16–20 June 2019; pp. 9376–9385.
21. Liu, Y.; Zhou, L.; Bai, X.; Huang, Y.F.; Gu, L.; Zhou, J.; Harada, T. Goal-oriented gaze estimation for zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–21 June 2021; pp. 3794–3803.

22. Xu, W.J.; Xian, Y.Q.; Wang, J.N.; Schiele, B.; Akata, Z. Attribute prototype network for zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21969–21980.
23. Wang, C.Q.; Min, S.B.; Chen, X.J.; Sun, X.Y.; Li, H.Q. Dual progressive prototype network for generalized zero-shot learning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2936–2948.
24. Chen, S.M.; Hong, Z.M.; Hou, W.J.; Xie, G.S.; Song, Y.B.; Zhao, J.; You, X.G.; Yan, S.C.; Shao, L. TransZero++: Cross attribute-guided transformer for zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 12844–12861. [[CrossRef](#)]
25. Li, C.W.; Xie, B.; Pei, Y. A RCS Periodicity extraction algorithm for ballistic target. In Proceedings of the International Conference on Image, Vision and Intelligent Systems, Changsha, China, 21–23 May 2021; pp. 1207–1216.
26. Wang, Z.H.; Luo, Y.; Li, K.M.; Yuan, H.; Zhang, Q. Micro-Doppler Parameters Extraction of Precession Cone-Shaped Targets Based on Rotating Antenna. *Remote Sens.* **2022**, *14*, 2549. [[CrossRef](#)]
27. Ren, K.; Du, L.; Lu, X.F.; Zhuo, Z.Y.; Li, L. Instantaneous frequency estimation based on modified Kalman filter for cone-shaped target. *Remote Sens.* **2020**, *12*, 2766. [[CrossRef](#)]
28. Wang, C.; Wen, S.L.; Ye, C.M. Three-dimensional reconstruction of space rotating target based on narrow-band radar networks. *J. Eng.* **2019**, *2019*, 6108–6112. [[CrossRef](#)]
29. Tang, W.B.; Yu, L.; Wei, Y.S.; Tong, P. Radar target recognition of ballistic missile in complex scene. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing, Chongqing, China, 11–13 December 2019; pp. 1–6.
30. Persico, A.R.; Ilioudis, C.V.; Clemente, C.; Soraghan, J.J. Novel classification algorithm for ballistic target based on HRRP frame. *IEEE Trans. Aerosp. Electron. Syst.* **2019**, *55*, 3168–3189. [[CrossRef](#)]
31. Persico, A.R.; Clemente, C.; Pallotta, L.; De Maio, A.; Soraghan, J. Micro-Doppler classification of ballistic threats using Krawtchouk moments. In Proceedings of the 2016 IEEE Radar Conference, Philadelphia, PA, USA, 2–6 May 2016; pp. 1–6.
32. Bai, X.R.; Wang, L.; Zhou, F.; Li, Y.G.; Hui, Y. Deep CNN for micromotion recognition of space targets. In Proceedings of the 2016 CIE International Conference on Radar, Guangzhou, China, 10–13 October 2016; pp. 1–5.
33. Wang, Y.Z.; Feng, C.Q.; Zhang, Y.S.; He, S.S. Space Precession Target Classification Based on Radar High-Resolution Range Profiles. *Int. J. Antennas Propag.* **2019**, *2019*, 8151620. [[CrossRef](#)]
34. Wang, S.R.; Li, M.M.; Yang, T.; Ai, X.; Liu, J.Q.; Andriulli, F.P.; Ding, D.Z. Cone-Shaped Space Target Inertia Characteristics Identification by Deep Learning with Compressed Dataset. *IEEE Trans. Antennas Propag.* **2022**, *70*, 5217–5226. [[CrossRef](#)]
35. Xu, G.G.; Yin, H.C.; Dong, C.Z. Micro-motion Forms Classification of Space Cone-shaped Target Based on Convolution Neural Network. *Appl. Computat. Electromagn. Soc. J.* **2020**, *35*, 64–71.
36. Wengrowski, E.; Purri, M.; Dana, K.; Huston, A. Deep CNNs as a method to classify rotating objects based on monostatic RCS. *IET Radar Sonar Navig.* **2019**, *13*, 1092–1100. [[CrossRef](#)]
37. Lee, J.; Kim, N.; Min, S.; Kim, J.; Jeong, D.; Seo, D. Space target classification improvement by generating micro-doppler signatures considering incident angle. *Sensors* **2022**, *22*, 1653. [[CrossRef](#)] [[PubMed](#)]
38. Wang, Y.Z.; Feng, C.Q.; Zhang, Y.S.; Ge, Q.C. Classification of space targets with micro-motion based on deep CNN. In Proceedings of the 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology, Harbin, China, 20–22 January 2019; pp. 557–561.
39. Zhang, Y.P.; Zhang, Q.; Kang, L.; Luo, Y.; Zhang, L. End-to-End Recognition of Similar Space Cone-Cylinder Targets Based on Complex-Valued Coordinate Attention Networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5106214. [[CrossRef](#)]
40. Han, L.X.; Feng, C.Q. Micro-Doppler-Based Space Target Recognition with a One-Dimensional Parallel Network. *Int. J. Antennas Propag.* **2020**, *2020*, 8013802. [[CrossRef](#)]
41. Li, R.; Wang, X.D.; Quan, W.; Zhang, G.L.; Xiang, Q. A staked discriminative auto-encoder based on center loss for radar target HRRP recognition. In Proceedings of the 2020 second International Conference on Artificial Intelligence Technologies and Application, Dalian, China, 21–23 August 2020; p. 012153.
42. Wang, X.D.; Li, R.; Wang, J.; Lei, L.; Song, Y.F. One-dimension hierarchical local receptive fields based extreme learning machine for radar target HRRP recognition. *Neurocomputing* **2020**, *418*, 314–325. [[CrossRef](#)]
43. Farhadi, A.; Endres, I.; Hoiem, D.; Forsyth, D. Describing objects by their attributes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1778–1785.
44. Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.S.; Dean, J. Zero-Shot Learning by Convex Combination of Semantic Embeddings. *arXiv* **2013**, arXiv:1312.5650.
45. Reed, S.; Akata, Z.; Lee, H.; Schiele, B. Learning Deep Representations of Fine-Grained Visual Descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 49–58.
46. Chen, L.; Zhang, H.W.; Xiao, J.; Liu, W.; Chang, S.F. Zero-Shot Visual Recognition using Semantics-Preserving Adversarial Embedding Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1043–1052.
47. Zhao, B.; Sun, X.W.; Yao, Y.; Wang, Y.Z. Zero-shot Learning via Shared-Reconstruction-Graph Pursuit. *arXiv* **2017**, arXiv:1711.07302.
48. Xu, Y.F.; Zhang, Q.M.; Zhang, J.; Tao, D.C. ViTAE: Vision Transformer Advanced by Exploring Intrinsic Inductive Bias. *arXiv* **2021**, arXiv:2106.03348.

49. Luo, C.J.; Zhan, J.F.; Xue, X.H.; Wang, L.; Ren, R.; Yang, Q. Cosine normalization: Using cosine similarity instead of dot product in neural networks. In Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018; pp. 382–391.
50. Huynh, D.; Elhamifar, E. Fine-grained generalized zero-shot learning via dense attribute-based attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4483–4493.
51. Zhu, Y.Z.; Xie, J.W.; Tang, Z.Q.; Peng, X.; Elgammal, A. Semantic-guided multi-attention localization for zero-shot learning. In Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 1–11.
52. Tian, X.D.; Bai, X.R.; Xue, R.H.; Qin, R.Y.; Zhou, F. Fusion Recognition of Space Targets with Micro-Motion. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 3116–3125. [[CrossRef](#)]
53. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.