remote sensing

MDPI

Article

Two-Stream Deep Fusion Network Based on VAE and CNN for Synthetic Aperture Radar Target Recognition

Lan Du *, Lu Li, Yuchen Guo, Yan Wang, Ke Ren and Jian Chen

National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China; luli92@stu.xidian.edu.cn (L.L.); ychguo@xidian.edu.cn (Y.G.); wangyanpx@163.com (Y.W.); renke@stu.xidian.edu.cn (K.R.); jianc@xidian.edu.cn (J.C.)

* Correspondence: dulan@mail.xidian.edu.cn

Abstract: Usually radar target recognition methods only use a single type of high-resolution radar signal, e.g., high-resolution range profile (HRRP) or synthetic aperture radar (SAR) images. In fact, in the SAR imaging procedure, we can simultaneously obtain both the HRRP data and the corresponding SAR image, as the information contained within them is not exactly the same. Although the information contained in the HRRP data and the SAR image are not exactly the same, both are important for radar target recognition. Therefore, in this paper, we propose a novel end-to-end two stream fusion network to make full use of the different characteristics obtained from modeling HRRP data and SAR images, respectively, for SAR target recognition. The proposed fusion network contains two separated streams in the feature extraction stage, one of which takes advantage of a variational auto-encoder (VAE) network to acquire the latent probabilistic distribution characteristic from the HRRP data, and the other uses a lightweight convolutional neural network, LightNet, to extract the 2D visual structure characteristics based on SAR images. Following the feature extraction stage, a fusion module is utilized to integrate the latent probabilistic distribution characteristic and the structure characteristic for the reflecting target information more comprehensively and sufficiently. The main contribution of the proposed method consists of two parts: (1) different characteristics from the HRRP data and the SAR image can be used effectively for SAR target recognition, and (2) an attention weight vector is used in the fusion module to adaptively integrate the different characteristics from the two sub-networks. The experimental results of our method on the HRRP data and SAR images of the MSTAR and civilian vehicle datasets obtained improvements of at least 0.96 and 2.16%, respectively, on recognition rates, compared with current SAR target recognition methods.

Keywords: target recognition; synthetic aperture radar; high-resolution range profile; fusion network; variational auto-encoder; convolutional neural network

Citation: Du, L.; Li, L.; Guo, Y.; Wang, Y.; Ren, K.; Chen, J. Two-Stream Deep Fusion Network Based on VAE and CNN for Synthetic Aperture Radar Target Recognition. *Remote Sens.* **2021**, *13*, 4021. https://doi.org/10.3390/rs13204021

Academic Editors: Ali Khenchaf and Jean-Christophe Cexus

Received: 17 August 2021
Accepted: 28 September 2021
Published: 9 October 2021

1. Introduction

Synthetic aperture radar (SAR) target recognition is a development of radar automatic target recognition (RATR) technology. Because of the all-weather, all-day and long-distance perception capabilities of SAR, SAR target recognition plays an important role in both military and civil fields [1–4]. SAR target recognition is urgently required, given the overwhelming amount of SAR data available, and the SAR target recognition has been a wide concern at home and abroad.

As a type of data widely used in RATR [5–10], high-resolution range profile (HRRP) data can be simultaneously obtained with the corresponding SAR image in the procedure of SAR imaging. HRRP data obtained from SAR echoes have been widely used for target recognition [11,12]. Figure 1 shows the relationship between HRRP data and the SAR image based on the classical range–Doppler algorithm (RDA) [13]. HRRP data is a 1D distribution of the radar cross section and can be obtained by the modulo operation after the range

compression of the received SAR echoes. HRRP target recognition receives widespread attention in the RATR community due to its relatively few complexities in signal acquisition [1–4]. A SAR image is the 2D image of the target derived by coherently processing high-range resolution radar echoes and conducting translational motion compensation by means of range cell migration correction (RCMC). The SAR images are easier and more intuitive to understand, being interpretable for human visual perception, as each pixel value reflects the surface microwave reflection intensity.

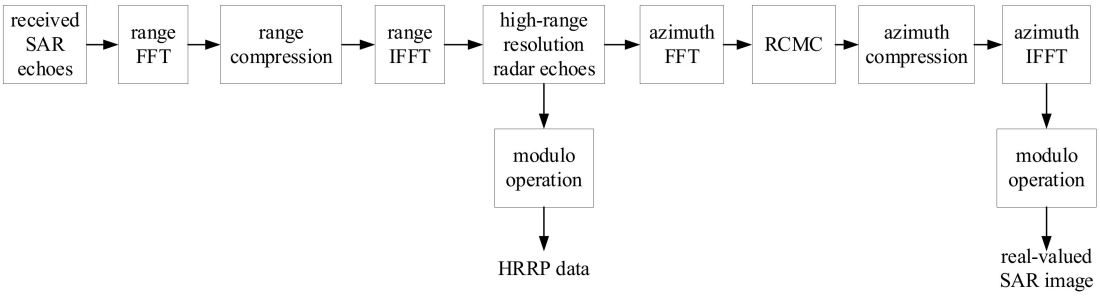


Figure 1. The data acquisition of HRRP data and real-valued SAR images from received SAR echoes.

Feature extraction is an important part of target recognition. The quality of the extracted features directly affects the performance of the target recognition. The development of HRRP data and SAR images in RATR has both gone through a process from the extraction of manual features to the extraction of depth features [1–4,14,15], which also leads to better RATR recognition performance. However, most of the existing RATR methods based on HRRP data and SAR images only use a single type of data. According to Figure 1, due to the different generation mechanisms, the information contained in the HRRP data and the SAR images is not exactly the same. Since the HRRP data and the SAR images can only represent the original SAR echoes from one aspect, using the above two data sources together can lead us to obtain a more complete information representation of the original SAR echoes. Modeling a complete interpretation using only unimodal data is theoretically insufficient. Therefore, to reveal more complete information, we underwent to formulate a novel framework to fuse the characteristics obtained from modeling HRRP data and the SAR image for radar target recognition. To the best of our knowledge, this is the first time that HRRP data and SAR images have been comprehensively utilized for radar target recognition.

In this paper, we propose an end-to-end two-stream fusion network. The first stream takes the HRRP data as the input, and draws support from the VAE, a deep probabilistic model, to effectively extract the latent probabilistic distribution features. The other stream takes the SAR image as its input. In this stream, a light weight CNN, LightNet, is utilized to extract 2D visual structure features. A fusion module with an attention mechanism is exploited to integrate the different characteristics extracted from two different signal types into a global space, to obtain a single, compact, comprehensive representation for radar target recognition that reflects target information more comprehensively and sufficiently. In the fusion module, the attention weight vector learned automatically is used to adaptively integrate the different characteristics, controlling the contribution of each feature to the overall output feature on a per-dimension basis, remarkably improving recognition performance. Finally, the fused feature is fed into a softmax layer to predict the classification results. More specifically, the main contributions of the proposed two-stream deep fusion network for target recognition are as follows:

1. Considering that both the SAR image and the corresponding HRRP data, in which the information contained are not exactly the same, can be simultaneously obtained in the procedure of SAR imaging, we apply two different sub-networks, VAE and LightNet, in the proposed deep fusion network to mine the different characteristics from the average profiles of the HRRP data and the SAR image, respectively. Through

joint utilization of these two types of characteristics, the target representation is more comprehensive and sufficient, which is beneficial for the target recognition task. Moreover, the proposed network is a unified framework which can be end-to-end joint optimized.

2. For the integration process on the latent feature of VAE and the structure feature of LightNet, a novel fusion module is developed in the proposed fusion network. The proposed fusion module takes advantage of the latent feature and the structure feature to automatically learn an attention weight. Then, the learned attention weight is used to adaptively integrate the latent feature and the structure feature. Compared with original concatenation operator, the proposed fusion module can achieve better recognition performance.

The rest of this paper is arranged as follows. Section 2 gives the related works of RATR based on HRRP data and SAR images. Section 3 introduces the novel two-stream fusion network. In Section 4, experiments based on the measured radar dataset and their corresponding analysis are presented to verify the target recognition performance of the proposed two-stream fusion method. Finally, the conclusions are presented in Section 5.

2. Related Work

2.1. Radar Target Recognition

Traditional radar target recognition methods are mostly based on manual feature extraction. These hand-crafted features are inappropriate if there is not sufficient prior knowledge on their application. Meanwhile, these features are mainly lower-level representations, e.g., textural features and local physical structural features, which cannot represent higher-level, abstract information.

Recently, deep learning has made progress by leaps and bounds in computer vision tasks due to its powerful representation capacity.

In HRRP target recognition, owing to the successful application of deep neural networks in various tasks, there are several deep neural networks have been developed on HRRP data. Part of the works focus on selecting suitable networks for HRRP recognition, such as stacked auto-encoder (SAE) [14], denoising auto-encoder (DAE) [5] and recurrent neural network (RNN) [16,17]. There are also some works which focus on how to use HRRP data reasonably, such as using the average profile of HRRP data and sequential HRRP data [18]. Nevertheless, those above-mentioned neural networks for HRRP recognition only gain the point estimations of latent features, which lack descriptions of the underlying probabilistic distribution. Considering the HRRP data do have the statistical distribution characteristics as [19–24] described, probabilistic statistical models are exploited to reveal the description of underlying probabilistic distribution, which can use prior information according to a solid theoretical basis, and an appropriate prior will enhance model performance. Meanwhile, probabilistic statistical models possess robustness and flexibility in modeling [25]. At present, several probabilistic statistical models have been developed to describe HRRP [23,26–28]. Nevertheless, traditional probabilistic models need to preset the distribution patterns of data, such as Gaussian distribution and Gamma distribution, which are relatively simple and have some limitations in their data fitting ability (the ability of fitting the original data distribution) [29]. In addition, since traditional probabilistic models are based on shallow architectures with simple linear mapping structures, they are only good at learning linear features. However, different from traditional probabilistic models, the VAE [6,30,31] introduces the neural network into probabilistic modelling. As we all know, neural networks stack nonlinear layers to form a deep structure. This nonlinear capability in VAE makes the data fitting more accurate, which can reduce the performance degradation caused by inaccurate data fitting. The deep structure of VAE can mine deep latent features of data with stronger feature separability. Because there is an explicit latent feature to represent the distribution characteristics of data in VAE, the latent variable is often directly used as the representational information of the sample for classification tasks, including HRRP target recognition [7,32,33], and has achieved good performance.

At present, VAE is the prevailing generative model. Meanwhile, the generative adversarial network (GAN) is also well known as a popular generative model. Although the VAE and the GAN both belong to generative models and they are usually mentioned at the same time, they are different in many aspects. In VAE, there is an explicit latent feature to represent the distribution characteristics of the data. Therefore, in the practical application of VAE, in addition to the common sample generation, the latent variable of VAE is often directly used as representational information of the sample for the classification and recognition tasks. However, restricted by the inherent mechanism of GAN, there is no explicit feature which can represent the distribution characteristics of data. The application of GAN focuses on the related fields of sample generation and transfer learning.

In the target recognition of the SAR images, auto-encoder (AE) [1,3] and the RBM [2], two widely used unsupervised deep neural network structures, are also employed and have better performance. Among deep neural networks, CNN has become the dominant deep learning approach, as in the VGG network [34], or ResNet. CNN architectures are usually comprised of multiple convolutional layers (followed with activation layers), pooling layers, and one or more fully connected layers. In CNNs, the local connection and weight share in the convolution operation, and the pool operation can effectively reduce the parameters and complexity, resulting in the invariance to translation and distortion which makes the learned features more robust [4]. Another advantage of the CNNs is that they can utilize convolution kernels to extract 2D visual structure information from the apparent to the abstract through layer-by-layer learning. This visual structure information plays a vital important role in image recognition [35–37].

In this paper, VAE and CNN are used as sub-networks for the HRRP data and the SAR image, respectively.

2.2. Information Fusion

In recent years, with the development of sensor technology, the diversity of information forms, the huge quantity of information, the complexity of information relations, and demand of timeliness, accuracy and reliability in information processing are unprecedented. Therefore, information fusion technology has been developed rapidly. Information fusion denotes the process of combining data from different sensors or information sources to obtain new or precise knowledge on physical quantities, events or situations [38].

According to the abstract level of information, information fusion methods can be divided into three categories: data-level fusion [39], feature-level fusion [40] and decision-level fusion [41]. Data-level and decision-level fusion are the two most easily implemented information fusion methods, but their performance improvements are also limited. Recently, it has also an important research topic to comprehensively and effectively use a variety of information of radar data, such as multi-temporal [42] and multi-view [43] data, to achieve better model performance. An inverse synthetic aperture radar (ISAR) target recognition method based on both range profile (RP) data and ISAR images was proposed, based on decision-level fusion of the classification results of RP data and ISAR images [44]. Feature-level fusion is the most effective method of information fusion, and it is often used as an effective means to improve performance in deep learning research. Several works focusing on image segmentation also use feature level fusion to fuse multi-level features [45–47]. However, these works fuse the features of the same data at different scales, while this paper fuses the features extracted from different data through their respective feature extraction networks.

3. Two-Stream Deep Fusion Network Based on VAE and CNN

The framework of the proposed two-stream deep fusion network for target recognition is depicted in Figure 2. As shown in Figure 2, the framework is briefly introduced as follows.

1. Data acquisition: as can be seen from Figure 1, the complex-valued high-range resolution radar echoes can be obtained after range compression of the receiving SAR echoes. Then, the HRRP data are obtained through the modulo operation. At the

same time, based on the complex-valued high-range resolution radar echoes, the complex-valued SAR image is obtained through azimuth focusing processing. Then, the commonly used real-valued SAR image for target recognition can be obtained by modulating the complex-valued SAR image.

2. VAE branch: based on the HRRP data, the average profile of the HRRP is obtained by preprocessing. Then, the average profile is fed into the VAE branch to acquire the latent probabilistic distribution as a representation of the target information.
3. LightNet branch: the other branch takes the SAR image as input and draws support from a lightweight convolutional architecture, LightNet, to extract the 2D visual structure information as another essential representation of the target information.
4. Fusion module: the fusion module is employed to integrate the distribution representation and the visual structure representation to reflect more comprehensive and sufficient information for target recognition. The fusion module merges the VAE branch and the LightNet branch into a unified framework which can be trained in an end-to-end manner.
5. Softmax classifier: finally, the integrated feature is fed into a usual softmax classifier to predict the category of target.

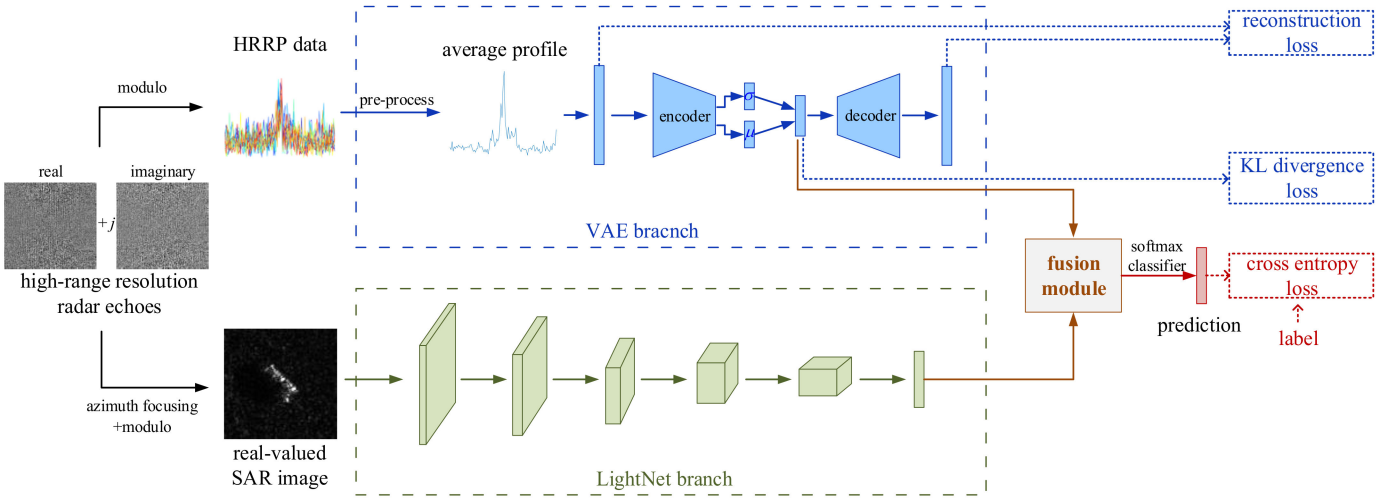


Figure 2. Framework of the proposed two-stream deep fusion network. Here the black solid lines and arrows represent the acquisition of inputs of the two sub-network branches, the blue solid lines and arrows represent the information flow in the VAE model, the green solid lines and arrows represent the information flow in the LightNet, the brown solid lines and arrows represent the information flow in the fusion module, and the red solid lines and arrows represent the final classifier. μ and σ represent the learned mean and standard deviation from the VAE encoder. The dotted line indicates the calculation of loss.

In Section 3.1, Section 3.2, Section 3.3, Section 3.4, Section 3.5, Section 3.6 some important components, including the acquisition data of the HRRP data and the real-valued SAR image from high-range resolution echoes, the VAE branch, the LightNet branch, the fusion module, the loss function and the training procedure, are introduced concretely.

3.1. Acquisition of the HRRP Data and the Real-Valued SAR Image from High-Range Resolution Echoes

Figure 1 in the Introduction gives the data acquisition procedure of the HRRP data and real-valued SAR image from received the SAR echoes based on RDA. The received SAR echoes are obtained from the radar-received signals through the dechirping and matched filters. The RDA SAR imaging algorithm can be divided into two steps: range focusing processing and azimuth focusing processing. The range focusing processing includes, in turn range fast Fourier transformation (FFT), range compression and range IFFT. Then,

the high-range resolution radar echoes can be obtained. The azimuth focusing processing includes, in turn, the azimuth FFT, RCMC, azimuth compressing and azimuth IFFT.

Based on the high-range resolution radar echoes, the HRRP data are obtained through the modulo operation. At the same time, based on the complex-valued high-range resolution radar echoes, the complex-valued SAR image is obtained through azimuth focusing processing. The azimuth focusing processing includes, in turn, the azimuth fast Fourier transformation (FFT), range cell migration correction (RCMC), azimuth compression and azimuth IFFT. Then, the commonly used real-valued SAR image for target recognition can be obtained by modulating the complex-valued SAR image. According to the introduction of the SAR imaging procedure, we can see that the complex SAR image is obtained using the high-range resolution radar echoes. Furthermore, given the complexity of the SAR image, the corresponding high-range resolution radar echoes and HRRP data also can be acquired [48,49].

Considering the mechanism inherent in the modulo operation, the modulo operation for generating HRRP data and the operation of the module for generating real-valued SAR images have different information loss characteristics. Therefore, although the HRRP data and the real-valued SAR images used in the proposed method keep a one-to-one correspondence, they cannot convert to each other anymore due to the operation of the module. In other words, the information contained in the HRRP data and the real-valued SAR images used in the proposed method are not exactly the same. The HRRP data and the SAR images can only represent the original high-range resolution radar echoes from one aspect each. Therefore, the features extracted from the HRRP data cannot be derived from the SAR images with certainty.

3.2. The VAE Branch

Before radar HRRP statistical modeling, there are some issues should be considered in practical application. The first is the time-shift sensitivity of HRRP. Centroid alignment [50] is commonly used as the time-shift compensation technique. We can eliminate amplitude-scale sensitivity through amplitude-scale normalization, such as L_2 normalization. Considering the target-aspect sensitivity [15,32], it has been demonstrated that the average profile has a smoother and more concise signal form than the single HRRP, and can better reflect the scattering property of the target in a given aspect-frame. From the perspective of signal processing, the average profile represents target's stable physical structure information in a frame [8,9,51]. One important characteristic of the average profile is that it can depress the speckle effect of HRRPs. Furthermore, the average profile also suppresses the impact of the noise spikes and the amplitude fluctuation property.

According to the literature [8,10,51], the definition of the average profile is

$$\mathbf{x}^{AP} = \left[\frac{1}{M}\sum_{i=1}^{M}\left|x^{P}{}_{i1}\right|, \frac{1}{M}\sum_{i=1}^{M}\left|x^{P}{}_{i2}\right|, \ldots, \frac{1}{M}\sum_{i=1}^{M}\left|x^{P}{}_{ir}\right| \right]^{T} = \frac{1}{M}\sum_{i=1}^{M}\left|\mathbf{x}^{P}{}_{i}\right| \tag{1}$$

where $\left\{\mathbf{x}^{P}{}_{i}\right\}_{i=1}^{M}$ is an HRRP frame, with the ith HRRP sample $\mathbf{x}^{P}{}_{i} = \left[x^{P}{}_{i1}, x^{P}{}_{i2}, \ldots, x^{P}{}_{ir}\right]^{T}$, and r is the dimension of HRRP samples.

The VAE holds that the sample space can be generated by the latent variable space, that is, sampling latent variables from a simpler latent variable space can generate the real samples within the sample space. The latent variable in VAE can describe the distribution characteristics of the data. The framework of VAE is illustrated in Figure 3. Given the observations $\left\{\mathbf{x}^{AP}{}_{n}\right\}_{n=1}^{N}$ with N samples, the VAE exploits an encoder model with input $\mathbf{x}^{AP}$ and outputs the mean, $\boldsymbol{\mu}$, and the standard deviation, $\boldsymbol{\sigma}$, of the latent variable, $\mathbf{z}$. Assuming the encoder model can be represented as f_{VAE_E} with parameter φ, which is also known as an inference model, $q_{\varphi}\left(\mathbf{z}\middle|\mathbf{x}^{AP}\right)$, the encoder of VAE can be formulated as follows:

$$\boldsymbol{\mu}, \boldsymbol{\sigma} = f_{VAE\_E}\left(\mathbf{x}^{AP}; \varphi\right). \tag{2}$$

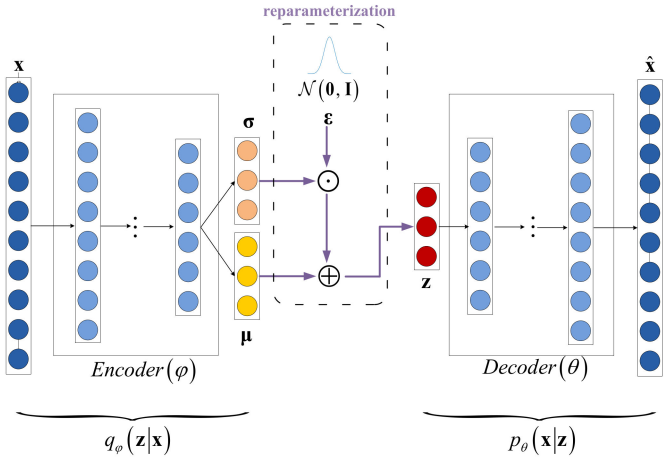


Figure 3. Architecture of the VAE with Gaussian distribution assumption.

Here, the reparametrization trick is adopted to sample from the posterior $\mathbf{z} \sim q_\varphi\left(\mathbf{z}|\mathbf{x}^{AP}\right)$ using the following:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0,\ \mathbf{I})$, and $\odot$ represents an element-wise product.

Then, with the latent variable $\mathbf{z}$ as the input, the decoder model f_{VAE_D} with parameter θ outputs the reconstruction sample, $\hat{\mathbf{x}}^{AP}$, which can be formulated as follows:

$$\hat{\mathbf{x}}^{AP} = f_{VAE\_E}(\mathbf{z}; \theta). \tag{4}$$

The decoder model is also known as a generative process with a probabilistic distribution: $p_\theta\left(\mathbf{x}^{AP}|\mathbf{z}\right)$.

The goal of the VAE model is to use the arbitrary distribution $q_\varphi\left(\mathbf{z}|\mathbf{x}^{AP}\right)$ to approximate the true posterior distribution $p_\theta\left(\mathbf{z}|\mathbf{x}^{AP}\right)$. Formally, as shown in Equation (5), the KL divergence is used to measure the similarity between $q_\varphi\left(\mathbf{z}|\mathbf{x}^{AP}\right)$ and $p_\theta\left(\mathbf{z}|\mathbf{x}^{AP}\right)p_\theta\left(\mathbf{z}|\mathbf{x}^{AP}\right)$, as follows:

$$KL\left(q_\varphi\left(\mathbf{z}\middle|\mathbf{x}^{AP}\right)\middle\|p_\theta\left(\mathbf{z}\middle|\mathbf{x}^{AP}\right)\right) = \log p_\theta\left(\mathbf{x}^{AP}\right) - LB\left(\theta, \varphi; \mathbf{x}^{AP}\right) \tag{5}$$

where

$$LB\left(\theta, \varphi; \mathbf{x}^{AP}\right) = E_{q_\varphi(\mathbf{z}|\mathbf{x}^{AP})}\left(\log p_\theta\left(\mathbf{x}^{AP}|\mathbf{z}\right)\right) - KL\left(q_\varphi\left(\mathbf{z}\middle|\mathbf{x}^{AP}\right)\|p_\theta(\mathbf{z})\right) \tag{6}$$

is the variational evidence lower bound (ELBO) [52,53].

For the given observations, $p_\theta\left(\mathbf{x}^{AP}\right)$ is a constant. Thus, minimizing the $KL\left(q_\varphi\left(\mathbf{z}|\mathbf{x}^{AP}\right)\|p_\theta\left(\mathbf{z}|\mathbf{x}^{AP}\right)\right)$ is equivalent to the ELBO maximization. Therefore, the loss of VAE on the data $\mathbf{x}^{AP}$ can be written as follows:

$$\begin{aligned} L_{VAE}\left(\theta, \varphi; \mathbf{x}^{AP}\right) &= LB\left(\theta, \varphi; \mathbf{x}^{AP}\right) \\ &= E_{q_\varphi(\mathbf{z}|\mathbf{x}^{AP})}\left(\log p_\theta\left(\mathbf{x}^{AP}|\mathbf{z}\right)\right) - KL\left(q_\varphi\left(\mathbf{z}|\mathbf{x}^{AP}\right)\|p_\theta(\mathbf{z})\right). \end{aligned} \tag{7}$$

In Equation (7), the first term can be regarded as reconstruction loss, which also can be written as follows:

$$E_{q_\varphi(\mathbf{z}|\mathbf{x}^{AP})}\left(\log p_\theta\left(\mathbf{x}^{AP}|\mathbf{z}\right)\right) = \left\|\mathbf{x}^{AP} - \hat{\mathbf{x}}^{AP}\right\|_2^2 \tag{8}$$

This teaches the decoder to reconstruct the data and suffers a cost if the output of decoder cannot reconstruct the data accurately. Usually, we can use a l_2-norm between the original data $\mathbf{x}^{AP}$ and the reconstructed data $\hat{\mathbf{x}}^{AP}$ as the reconstruction loss. The second term is the *KL* divergence between the encoder's distribution $q_\varphi\left(\mathbf{z}|\mathbf{x}^{AP}\right)$ and the prior $p_\theta(\mathbf{z})$. Typically, if we let the prior over the latent variables be the centered isotropic multivariate Gaussian $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \mathbf{I})$, the *KL* divergence in Equation (7) can be computed as follows:

$$KL\left(q_\varphi\left(\mathbf{z}\middle|\mathbf{x}^{AP}\right)\|p_\theta(\mathbf{z})\right) = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log(\sigma_j)^2 - (\mu_j)^2 - (\sigma_j)^2\right) \tag{9}$$

where μ_j and σ_j represent the jth element in the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, respectively, and J denotes the dimensionality.

Then, Equation (7) can be rewritten as follows:

$$L_{VAE}\left(\theta, \varphi; \mathbf{x}^{AP}\right) = \left\|\mathbf{x}^{AP} - \hat{\mathbf{x}}^{AP}\right\|_2^2 + \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\left((\sigma_j)^2\right) - (\mu_j)^2 - (\sigma_j)^2\right) \tag{10}$$

where $\|\cdot\|_2$ denotes the l_2-norm.

In practice, the encoder model is implemented with a three-layer, fully connected neural network. The units in the encoder model are 512, 256, and 128 respectively. Moreover, the decoder model is also implemented with a three-layer, fully connected neural network. The units in the decoder model are 128, 256 and 512 respectively. The dimensions of the latent variable $\mathbf{z}$, the mean $\boldsymbol{\mu}$ and the standard deviation $\boldsymbol{\sigma}$ are set to 50.

3.3. The LightNet Branch

Among deep neural networks, CNNs have made remarkable progress due to their characteristics of local connection and weight sharing. CNNs take advantage of convolution kernels to extract 2D visual structure information through layer-by-layer processing. Many excellent convolutional network architectures, such as VGG and ResNet, have come to dominate many fields. Nevertheless, considering the limited data volume, these above-mentioned networks still have a larger number of parameters for the task of SAR target recognition. Therefore, we applied a lightweight CNN, called LightNet, which has very few parameters and can achieve approximate performance.

The LightNet architecture is mainly comprised of convolution layers and pooling layers. Following each convolution layer, there is a rectified linear unit (ReLU) as an activation function and a batch normalization layer which allows the network to use much higher learning rates and be less careful about initialization [54]. The architecture of the LightNet is shown in Table 1. In the LightNet, there are only 5 convolutional layers. The kernel size in the first convolutional layer is 11×11, which is a larger kernel size, for gaining a larger receptive field. In the following three convolutional layers, the kernel sizes are 5×5. Considering that the fully connected layer in the original LightNet, which is usually used to transform feature maps to a feature vector at the final position in the network, has many parameters, we use a convolutional layer with a 3×3 kernel and no padding to replace a common fully connected layer to generate a feature vector from the feature maps. The convolutional layer has fewer parameters than the fully connected layer. Compared with global pooling, the convolutional layer can not only learn more abstract features but also adjust the dimensions of the feature vector.

Table 1. The architecture of the LightNet used in our method.

Input	Operator	Kernel Size	Number of Channels	Strides
$128 \times 128 \times 1$	Convolution	11	16	2
$62 \times 62 \times 16$	Pooling	2	16	2
$31 \times 31 \times 16$	Convolution	5	32	1
$27 \times 27 \times 32$	Pooling	2	32	2
$14 \times 14 \times 32$	Convolution	5	64	1
$10 \times 10 \times 64$	Pooling	2	64	2
$5 \times 5 \times 64$	Convolution	5	128	1
$3 \times 3 \times 128$	Convolution	3	100	1

The LightNet network considers the SAR image, $\mathbf{x}^I$, as an input to extract the 2D visual structure information, $\mathbf{m}$, as another essential representation of the target information. Assuming f_{LNet} represents the LightNet with parameter ψ_{LNet}, then the LightNet branch can be formulated as follows:

$$\mathbf{m} = f_{LNet}\left(\mathbf{x}^I; \psi_{LNet}\right) \tag{11}$$

3.4. Fusion Module

In the feature extraction stage, a VAE model is employed on the HRRP data to extract the latent probabilistic distribution information as a feature, and a lightweight LightNet is used on the SAR image to extract the structure features. In the neural network framework, the most common feature fusion approaches are the concatenation operation and element-wise addition. The concatenation operation combines multiple original features according to the feature dimensions to generate a fused feature, and the dimension of the fused features is equal to the sum of the original feature dimensions. Although it is simple to realize, the dimension of the fused feature is relatively high, which brings a greater pressure on the subsequent classifiers, including the increase in the number of parameters and the cost of optimizing the parameters. The element-wise addition is also a common feature fusion method. Based on the element-wise addition, the fusion feature is obtained by adding the original features, element by element, which keeps the dimension consistent with the original features and requires smaller parameters on the subsequent classifiers than the concatenation operation. In essence, element-wise addition assumes that the importance of different features is the same.

To reflect the target information more comprehensively and sufficiently, a novel fusion module was exploited to integrate the latent feature obtained from VAE and the structure feature obtained from LightNet, which can also merge the two streams into a unified framework with end-to-end joint optimization. The proposed fusion module is a further extension on the element-wise addition inspired by the gated recurrent unit (GRU) [55]. On the one hand, we use an attention weight vector, not a single value, to integrate the different features. More clearly, in the feature fusion, we no longer think that each dimension in a feature vector shares the same weight, but each dimension of the feature vector has its own weight coefficient. The influence of features that contribute more to the target task on the fusion features is increased by considering the differences in the importance of each feature more carefully. Likewise, the influence of features that contribute less to the target task on fusion features is weakened. On the other hand, compared with traditional, empirically set weight values, the attention weight vector is learned automatically according the target task, which can perform an adaptive adjustment of feature weights with the samples and categories.

Figure 4 shows the flowchart of the fusion module. At first, the latent feature, $\mathbf{z}$, and the structure feature, $\mathbf{m}$, are fed into fully connected layers, respectively, to generate the features $\widetilde{\mathbf{Z}} \in \mathbb{R}^{d\times 1}$ and $\widetilde{\mathbf{M}} \in \mathbb{R}^{d\times 1}$:

$$\begin{aligned} \widetilde{\mathbf{Z}} &= \mathrm{ReLU}(\mathbf{W}_Z \cdot \mathbf{z}) \\ \widetilde{\mathbf{M}} &= \mathrm{ReLU}(\mathbf{W}_M \cdot \mathbf{m}) \end{aligned} \tag{12}$$

where features $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{M}}$ have the consistent dimension d which was set to 50 for the experiments. $\mathrm{ReLU}(\cdot)$ denotes the ReLU activation operation, and $\mathbf{W}_Z$ and $\mathbf{W}_M$ are the parameters in the fully connected layers, respectively. Here, the fully connected layers are applied not only to further map two features into a global space, but also to make the dimension and order contain consistent correspondence for subsequent element-wise addition, i.e., the relationship between the ith feature in $\widetilde{\mathbf{Z}}$ and the ith feature in $\widetilde{\mathbf{M}}$ is a one-to-one correspondence relationship.

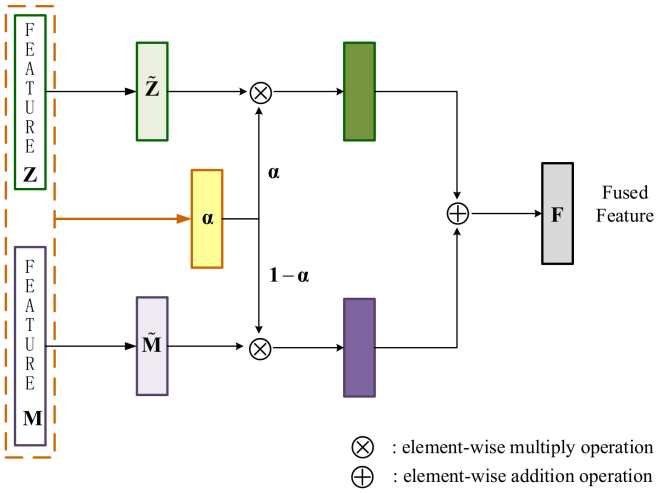


Figure 4. Flowchart of the proposed fusion module.

Then, the latent feature $\mathbf{z}$ and the structure feature $\mathbf{m}$ are concatenated into a long feature vector, and a fully connected layer is used on the long feature vector to learn the attention vector $\boldsymbol{\alpha} \in \mathbb{R}^{d\times 1}$:

$$\boldsymbol{\alpha} = sigmoid(\mathbf{W}_\alpha \cdot [\mathbf{z}, \mathbf{m}]) \tag{13}$$

where $sigmoid(\cdot)$ denotes the sigmoid activation operation, and $\mathbf{W}_\alpha$ denotes the parameter. Drawing support from the sigmoid activation, the value in the attention vector is in the range $[0, 1]$. Here, the attention mechanism is derived from the selective attention behavior of the human brain when processing information. The human brain scans the total information quickly to get the focus area, and then invests more attention resources in this area to obtain more detailed information of the target task, while suppressing other useless information. This method has greatly improved the means of screening high-value information from a large quantity of information. Similar to the selective attention mechanism of human beings, the core goal of the attention mechanism we used was to select the information that as most critical to the current task from a large quantity of information. Therefore, a fully connected layer with activation was used to simulate the neurons in the human brain. The input of the fully connected layer was all of the sample information, i.e., all the features of the sample. By using the fully connected layer to sense all the information, we could determine the focus area/features, and then invest more attention on the focus features while suppressing other useless information. That is to say, we can know where to focus and the degree to which to focus from the output of the fully connected layer. Therefore, the output of the fully connected layer is called attention weight vector.

Finally, the attention vector $\boldsymbol{\alpha}$ was used as a weight to sum the $\widetilde{\mathbf{Z}}$ and $\widetilde{\mathbf{M}}$. Due to the value of $\boldsymbol{\alpha}$ being in range $[0, 1]$, the value in $1 - \boldsymbol{\alpha}$ is also in range $[0, 1]$. The attention vector can be regarded as a weight vector which controls the contribution of the feature $\widetilde{\mathbf{Z}}$ to the overall output of the unit. In contrast, considering the weight normalization, the weight of feature $\widetilde{\mathbf{M}}$ can be directly obtained by the operation $1 - \boldsymbol{\alpha}$ without an extra learning process. More concretely, the attention vector $\boldsymbol{\alpha}$ is element-wise multiplied with the feature $\widetilde{\mathbf{Z}}$, and the vector $1 - \boldsymbol{\alpha}$ is element-wise multiplied with the feature $\widetilde{\mathbf{M}}$, and then the element-wise sum operation is used to integrate these two features:

$$\mathbf{F} = \boldsymbol{\alpha} \otimes \widetilde{\mathbf{Z}} + (1 - \boldsymbol{\alpha}) \otimes \widetilde{\mathbf{M}} \tag{14}$$

where $\otimes$ represents the element-wise multiply operation, 1 is a vector whose elements are all valued one, and $\mathbf{F}$ represents the fusion feature.

Assuming f_{fusion} represents the overall fusion module with the parameter ψ_{fusion}, then the fusion module can be summarized as follows:

$$\mathbf{F} = f_{fusion}\left(\mathbf{z}, \mathbf{m}; \psi_{fusion}\right) \tag{15}$$

3.5. Loss Function

Following the feature extraction stage and the fusion module, the fusion feature $\mathbf{F}$ was fed into a softmax layer to predict the classification results $\{\hat{\mathbf{y}}_n\}_{n=1}^{N}$, which can be formulated as follows:

$$\hat{\mathbf{y}} = fc(\mathbf{F}; \psi_c) \tag{16}$$

where fc represents a usual softmax classifier with parameter ψ_c.

The supervised constraint ensures that the prediction label $\hat{\mathbf{y}}_n$ is closed to the true label $\mathbf{y}_n$ via the cross-entropy loss function, as follows:

$$L_{label}(\mathbf{y}_n, \hat{\mathbf{y}}_n) = -\sum_{k=1}^{K} y_n^k \log\left(\hat{y}_n^k\right) \tag{17}$$

where K represents the number of classes.

Therefore, the total loss function of the proposed deep fusion network for target recognition is a combination of L_{label} and L_{VAE} (described in formula (10)) as follows:

$$L_{total} = L_{label} + L_{VAE} \tag{18}$$

3.6. Training Procedure

Based on the total loss function, L_{total}, the backpropagation algorithm where we used stochastic gradient descent (SGD), was used for the proposed network for end-to-end joint optimization. The total training procedure of the proposed network is outlined in Algorithm 1.

Algorithm 1. Training Procedure of the Proposed Network

1. Set the architecture of the proposed network, including the number of fully connected layers, the units in each fully connected layer, the size of convolutional kernels, the strides and the number of channels, and so on.
2. Initialize the network parameters φ, θ, ψ_{LNet}, ψ_{fusion} and ψ_c.
3. **while** not converged **do**
4. Randomly sample a mini-batch $\{\mathbf{X}_b\}_{b=1}^{B}$ and its corresponding label $\{\mathbf{y}_b\}_{b=1}^{B}$ from the whole dataset.
5. Based on each data $\mathbf{X}_b$ in the mini-batch, generate the average profile ${\mathbf{x}_b}^{AP}$ and the SAR image ${\mathbf{x}_b}^{I}$.
6. Sample random noise $\{\boldsymbol{\varepsilon}_n\}_{n=1}^{N_1}$ from standard Gaussian distribution for re-parameterization.
7. With ${\mathbf{x}_b}^{AP}$ as input, generate the latent distribution representation $\mathbf{z}$ using Equations (2) and (3), and then generate the reconstruction $\hat{\mathbf{x}}^{AP}$ based on Equation (4).
8. With ${\mathbf{x}_b}^{I}$ as input, generate the structure information $\mathbf{m}$ using Equation (11).
9. Based on Equation (15), generate integrated feature $\mathbf{F}$.
10. Based on integrated feature $\mathbf{F}$, obtain prediction $\hat{\mathbf{y}}$ with Equation (16).
11. Compute the total loss L_{total}.
12. Update network parameters φ, θ, ψ_{LNet}, ψ_{fusion} and ψ_c via SGD on the total loss L_{total}.
13. **end while**

4. Results

4.1. Experimental Data Description

Experiments were carried out based on the HRRP data and SAR images of the moving and stationary target acquisition and recognition (MSTAR) dataset, which was collected through a HH polarization mode SAR sensor working in the X-band with 0.3 × 0.3 m resolution in spotlight mode [56]. The MSTAR dataset, a measured benchmark dataset, is widely used for evaluating SAR target recognition performance. The MSTAR dataset includes ten different ground military targets, i.e., BMP2 (tank), BTR70 (armored vehicle), T72 (tank), BTR60 (armored vehicle), 2S1 (cannon), BRDM (truck), D7 (bulldozer), T62 (tank), ZIL131 (truck) and ZSU234 (cannon). Among them, BMP2 and T72 have variants in the test stage. The depression angles of the samples for each target category are 15° and 17°, and the aspect angles cover a range from 0° to 360°. Referring to the existing literature, this paper focuses on two experimental scenarios: three-target data SAR target recognition and ten-target data SAR target recognition. The specific details of the experimental data setting on the above-mentioned two experimental scenarios are listed in Tables 2 and 3, respectively. Optical image examples of the ten different targets are shown in Figure 5, and the corresponding SAR image examples are listed in Figure 6.

Table 2. Details of Training and Test Samples for the Three-Target Dataset.

Dataset	BMP2			BTR70		T72	
	C21	9566	9563	C71	132	S7	812
Training samples (17°)	233	0	0	233	232	0	0
Test samples (15°)	196	196	195	196	196	191	195

Table 3. Details of Training and Test Samples for the Ten-Target Dataset.

Dataset	BMP2			BTR70	T72		BTR60	2S1	BRDM	D7	T62	ZIL 131	ZSU 234
Training samples (17°)		233 (C21)		233	232 (132)		255	299	298	299	299	299	299
Test samples (15°)	196 (C21)	196 (9566)	195 (9563)	196	196 (132)	191 (S7)	195 (812)	195	274	274	274	273	274	274

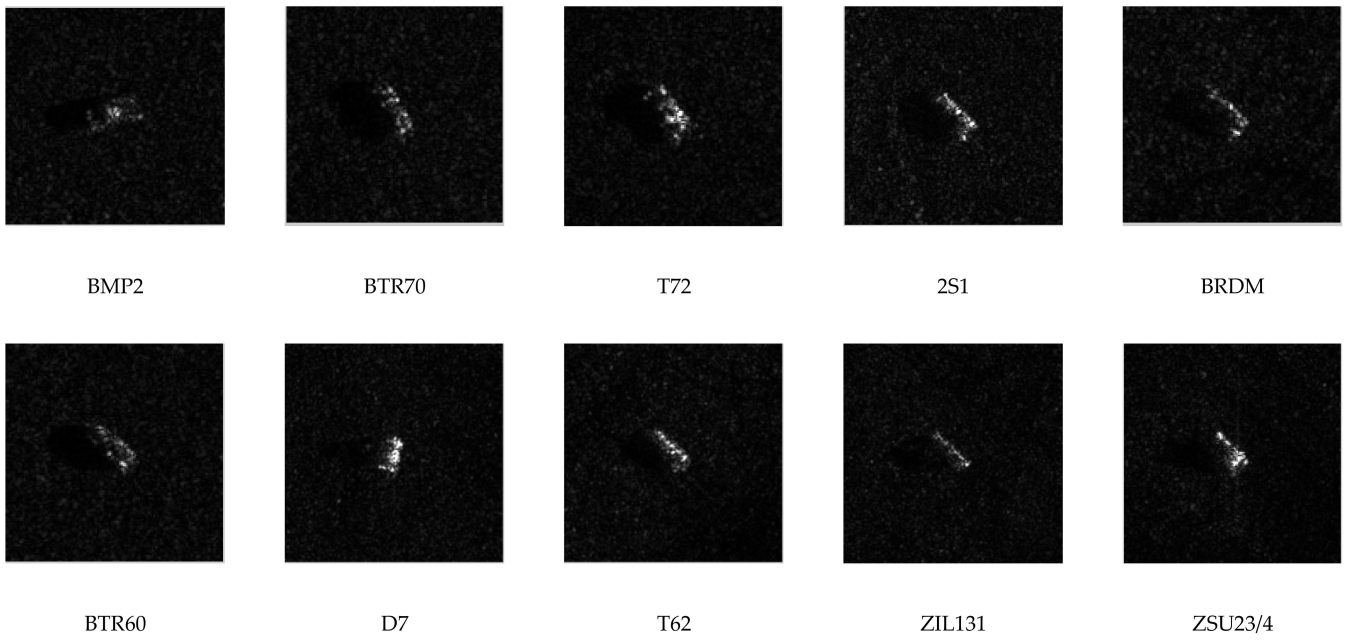


Figure 5. SAR image examples of ten different targets in the MSTAR dataset.

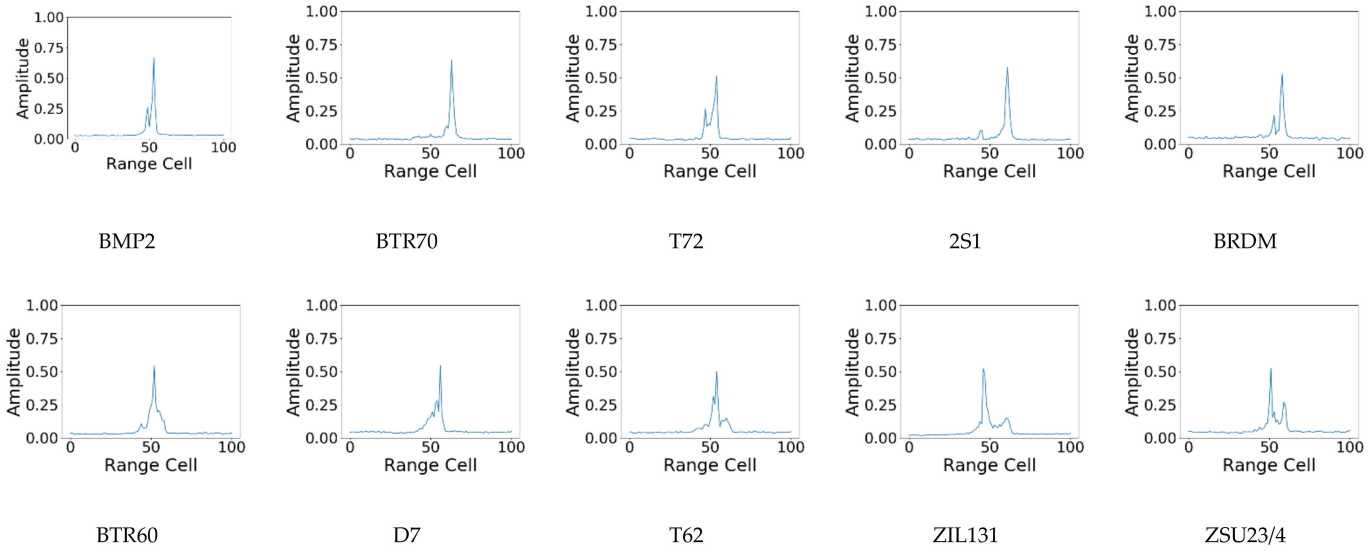


Figure 6. The average profile examples of the generated HRRP data of the ten different targets in the MSTAR dataset.

For the MSTAR data, we use the complex-valued SAR images provided by the U.S. Defense Advanced Research Projects Agency and the U.S. Air Force Research Laboratory to get the high-range resolution radar echoes in reverse without information loss, in accordance with the reference [11]. And then, based on the high-range resolution radar echoes, the HRRP data can be generated, as shown in Figure 1. The average profile

examples of the generated HRRP data of the ten different targets are listed in Figure 6. The real-valued SAR image was directly obtained by a modulo operation on the complex-valued SAR images.

4.2. Evaluation Criteria

For the quantitative analysis, we use two widely used criteria, namely, the overall accuracy and the average accuracy, as the evaluation criteria to evaluate target recognition performance.

$$overall\ accuracy = \frac{\sum_{i=1}^{N_C} \mathrm{Tr}_i}{\sum_{i=1}^{N_C} Q_i} \tag{19}$$

$$average\ accuracy = \frac{1}{N_C} \sum_{i=1}^{N_C} \frac{\mathrm{Tr}_i}{Q_i} \tag{20}$$

where Tr_i represents the number of test samples recognized correctly in class i, Q_i represents the total number of test samples in class i, and N_C represents the number of classes.

The higher the values of the overall accuracy and the average accuracy, the better the performance of target recognition method.

4.3. Three-Target MSTAR Data Experiments

In this section, we discuss the effectiveness of the proposed method on the three-target MSTAR data. We gave the confusion matrix of the proposed deep fusion network on three-target MSTAR data in Table 4. The confusion matrix is a widely used performance evaluation method for target recognition. In a confusion matrix, each row represents the actual category, while each column is the predicted category, and the elements denote the probabilities that the targets are recognized as a certain class. In particular, the elements on the diagonal represent the recognition accuracy. From Table 4, it is easy to see that the accuracy on BTR70 was 0.9898, the accuracy on T72 was 0.9880, and the accuracy on BMP2 was 0.9642, which shows the proposed method has better recognition performance.

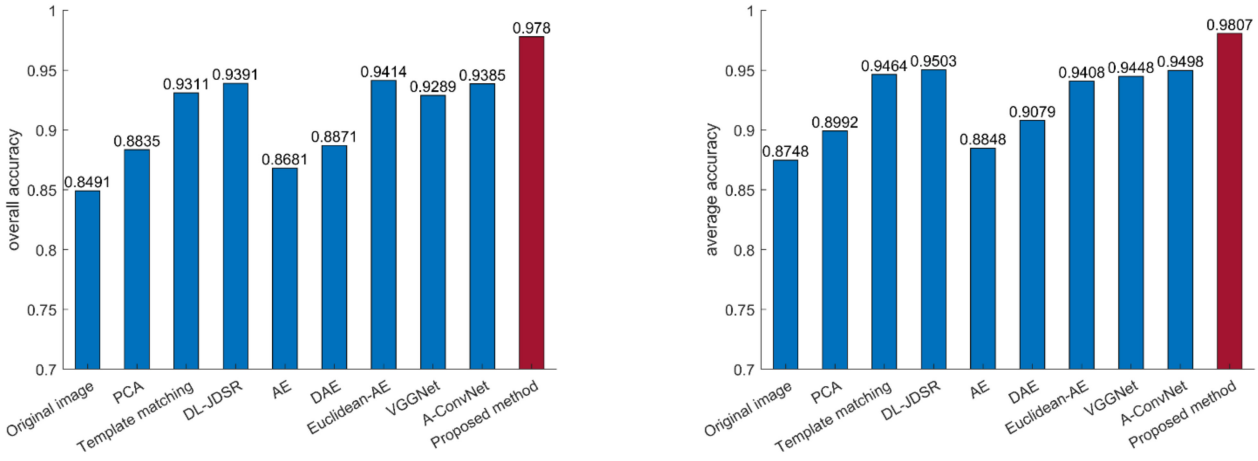
Table 4. Confusion Matrix of the Proposed Method on Three-Target Data.

Type	BMP2	BTR70	T72
BMP2	0.9642	0.0034	0.0324
BTR70	0.0102	0.9898	0
T72	0.0103	0.0017	0.9880

In order to further validate the efficiency of the proposed method, we compared the proposed method with some traditional SAR target recognition methods, i.e., directly applying the amplitude feature of the original SAR images, principal component analysis (PCA), the template matching method, dictionary learning and JDSR (DL-JDSR) [57], sparse representation in the frequency domain (SRC-FT) [58], and Riemannian manifolds [59]. Moreover, the proposed method was compared with other deep learning-based target recognition methods without data augmentation, as seen in Table 5. The compared deep learning-based target recognition methods include the original auto-encoder (AE), denoising AE (DAE), linear SVM, the Euclidean distance restricted AE [3] (Euclidean-AE), the VGG convolutional neural network (VGGNet), A-ConvNets [60], the early feature fusion of a model-based geometric hashing (MBGH) approach and a CNN approach (MBGH+CNN with EFF) [61], compact convolutional autoencoder (CCAE) [62], ResNet-18 [63], ResNet-34 [63] and DenseNet [64]. Figure 7 shows the intuitional accuracy results of the proposed method and the above-mentioned compared methods. Table 5 lists their detailed accuracies with the three-target MSTAR data and their overall and average accuracies.

Table 5. Detailed Accuracies of Different Types on The Three-Target Data via Some SAR Recognition Methods.

	BMP2	BTR70	T72	Overall Accuracy	Average Accuracy
proposed method	**0.9642**	0.9898	**0.9880**	**0.9780**	**0.9807**
original image	0.7325	0.9643	0.9278	0.8491	0.8748
PCA	0.8330	0.9541	0.9106	0.8835	0.8992
Template matching	0.9148	**1**	0.9244	0.9311	0.9464
DL-JDSR [65]	0.9301	0.9898	0.9312	0.9391	0.9503
AE	0.8756	0.9439	0.8351	0.8681	0.8848
DAE	0.7922	0.9796	0.9519	0.8871	0.9079
Euclidean-AE [3]	0.9421	0.9388	0.9416	0.9414	0.9408
VGGNet	0.8859	**1**	0.9485	0.9289	0.9448
A-ConvNets	0.9199	0.9898	0.9399	0.9385	0.9498
ResNet-18	0.9642	1	0.9485	0.9626	0.9709
ResNet-34	0.9676	0.9847	0.9622	0.9678	0.9715
DenseNet	0.8756	0.9592	0.8625	0.8821	0.8991
SRC-FT [58]	0.9625	1	0.9519	0.9631	0.9715
Riemannian manifolds [59]	0.9574	0.9847	0.9570	0.9610	0.9664
CCAE [62]	0.9523	**1**	0.9742	0.9684	0.9755
MBGH+CNN with EFF	0.9387	0.9643	0.9313	0.9389	0.9448



(**a**) (**b**)

Figure 7. Three-target accuracies obtained by different SAR target recognition methods. (**a**) overall accuracy; (**b**) average accuracy.

From Figure 7 we can clearly see that, compared with the original image method and the PCA, template matching, DL-JDSR, SRC-FT and Riemannian manifold methods, our proposed method performs better on both overall accuracy and average accuracy. The proposed method also yields higher overall and higher average accuracy than the compared deep learning methods, i.e., AE, DAE, Euclidean-AE, VGGNet, A-ConvNet, MBGH+CNN with EFF, ResNet-18, ResNet-34 and DenseNet. As shown in Table 5, for the BMP2 and the T72 types with variants in the test stage, the accuracies of the proposed method attained 0.9642 and 0.9880, respectively, which outperformed all other compared target recognition methods. For the BTR70 type, which does not contain variants, the template matching method and VGGNet could correctly recognize all test samples, at the same time, the accuracies of the proposed method, DL-JDSR and the A-ConvNet were also 0.9898, which is very close to 1. In terms of overall accuracy and average accuracy in Table 5, two comprehensive evaluation criteria, we can see that the proposed method is at least 0.96% and 0.52% higher, respectively than other compared methods.

4.4. Ten-Target MSTAR Data Experiments

In this section, we evaluate the target recognition performance of the proposed method with the ten-target MSTAR data. Similar to the Section III-C, the confusion matrix is shown at first in Table 6. From the Table 6 we can see that the accuracy of all target types, except T72, was over 0.97. The best accuracy is shown in ZIL131, where the test samples were all correctly classified. The accuracies on BTR60, D7 and ZSU234 were close to 1, and the worst accuracy was over 0.94.

Table 6. Confusion Matrix of the Proposed Method on Ten-Target Data.

Type	BMP2	BTR70	T72	BTR60	2S1	BRDM	D7	T62	ZIL131	ZSU234
BMP2	0.9710	0.0034	0.0239	0.0017	0	0	0	0	0	0
BTR70	0.0051	0.9949	0	0	0	0	0	0	0	0
T72	0.0275	0.0034	0.9433	0	0.0069	0	0	0.0034	0.0069	0.0086
BTR60	0	0	0.0051	0.9846	0.0103	0	0	0	0	0
2S1	0.0109	0.0036	0.0073	0	0.9745	0	0	0	0.0036	0
BRDM	0.0109	0	0	0.0036	0	0.9818	0	0	0.0036	0
D7	0	0	0	0	0	0	0.9927	0	0.0073	0
T62	0	0	0.0183	0	0	0	0	0.9707	0.0110	0
ZIL131	0	0	0	0	0	0	0	0	1	0
ZSU234	0	0	0	0	0	0	0	0	0.0036	0.9964

We compare the performance of the proposed method with the original image, PCA, template matching, DL-JDSR, AE, DAE, Euclidean-AE, VGG network, A-ConvNets, MBGH+CNN with EFF, ResNet-18, ResNet-34 and DenseNet methods in Figure 8 and Table 7.

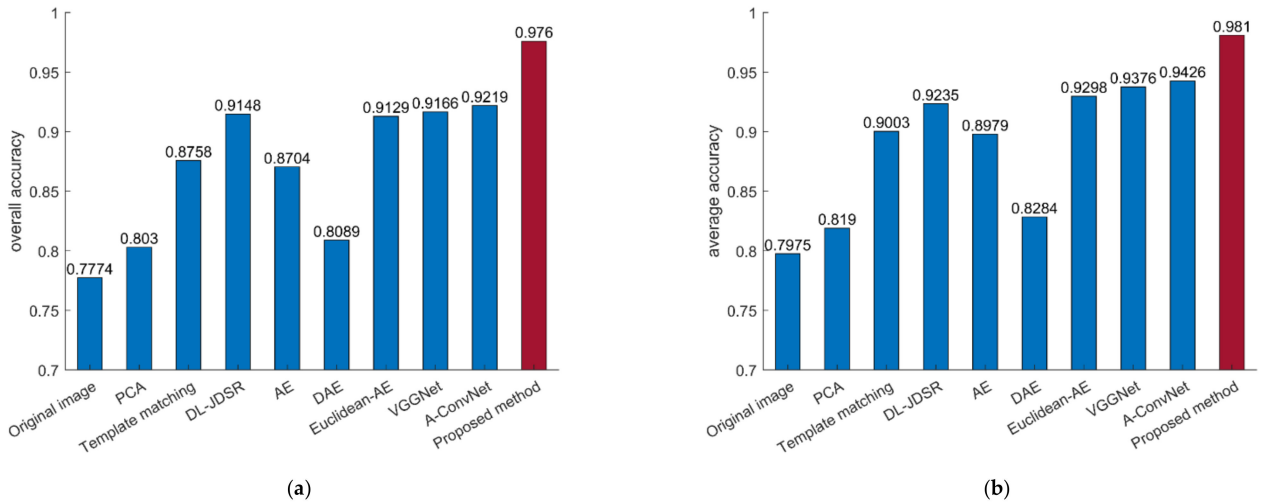


Figure 8. Ten-target accuracies obtained by different SAR target recognition methods. (**a**) Overall accuracy; (**b**) average accuracy.

As shown in Figure 8 and Table 7, our proposed method outperforms all the other compared methods. Especially, for the first nine types, i.e., BMP2, BTR70, T72, BTR60, 2S1, BRDM, D7 and T62, the proposed method yielded the highest accuracy. For the ZSU234 type, the A-ConvNet method has the highest accuracy and the proposed method followed closely, with an accuracy of 0.9964. In terms of overall accuracy, we can see that the proposed method is at least 4% higher than the other compared methods. The proposed method is about 4% higher in terms of average accuracy.

Table 7. Detailed Accuracies of Different Types in the Ten-Target Data via Some SAR Recognition Methods.

	BMP2	BTR70	T72	BTR60	2S1	BRDM	D7	T62	ZIL131	ZSU234	Overall Accuracy	Average Accuracy
Proposed method	**0.9710**	**0.9949**	**0.9433**	**0.9846**	**0.9745**	**0.9818**	**0.9927**	**0.9707**	**1**	0.9964	**0.9760**	**0.9810**
Original image	0.6899	0.8673	0.7131	0.7897	0.4453	0.9307	0.8905	0.7802	0.9124	0.9562	0.7774	0.7975
PCA	0.7070	0.8520	0.7715	0.8051	0.6971	0.7920	0.9598	0.8645	0.7956	0.9453	0.8030	0.8190
Template matching	0.8637	0.9235	0.6993	0.9179	0.8577	0.8869	0.9818	0.9670	0.9307	0.9745	0.8758	0.9003
DL-JDSR [65]	0.8876	0.9388	0.8625	0.8821	0.8905	0.9161	0.9854	0.9670	0.9234	0.9818	0.9148	0.9235
AE	0.8245	0.9082	0.6735	0.8718	0.9161	0.9562	0.9708	0.9451	0.9234	0.9891	0.8704	0.8979
DAE	0.7155	0.8980	0.7371	0.7436	0.5657	0.9599	0.9088	0.8498	0.9380	0.9672	0.8089	0.8284
Euclidean AE [3]	0.8790	0.9286	0.7955	0.9179	0.9380	0.9672	0.9891	0.9414	0.9453	0.9964	0.9129	0.9298
VGGNet	0.7683	0.9745	0.8872	0.9270	0.9818	0.9964	0.9780	0.9891	0.9854	0.8883	0.9166	0.9376
A-ConvNets	0.8961	0.9745	0.7887	0.9641	0.9197	0.9818	0.9526	0.9597	0.9891	**1**	0.9219	0.9426
ResNet-18	0.9216	0.9541	0.8488	0.8923	0.9562	0.9161	0.9051	0.9341	0.8869	0.9562	0.9107	0.9171
ResNet-34	0.8842	0.9286	0.8677	0.7795	0.9343	0.8759	0.9011	0.9234	0.9161	0.9380	0.8932	0.8949
DenseNet	0.9438	0.9745	0.9158	0.8410	0.9453	0.9526	0.9194	0.9745	0.9234	0.9708	0.9360	0.9361
MBGH+CNN with EFF	0.8518	0.8929	0.9553	0.8307	0.8723	0.9015	0.8864	0.8942	0.9343	0.8942	0.8876	0.8914

4.5. Model Analysis

4.5.1. Ablation Study

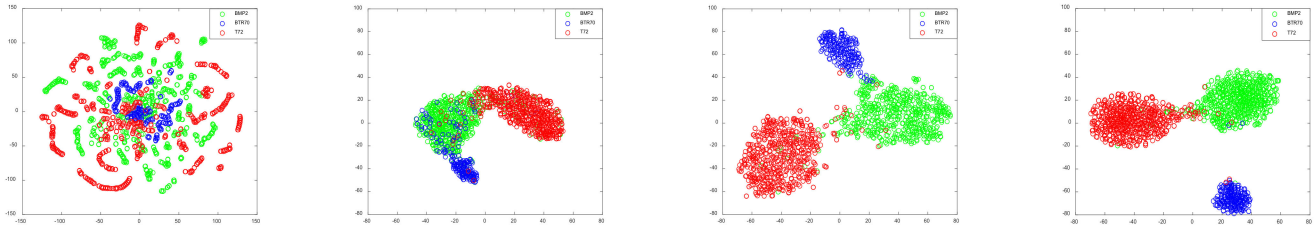
In order to gain a better understanding of the network's behavior and prove that the fusion of HRRP data and SAR images is beneficial to SAR target recognition, an ablation study is always adopted to see how each component affects the performance where one or more certain components of the network are removed or replaced. Therefore, in this sub-section, several controlled experiments are designed. Except for certain examined components, the rest of settings remain consistent. The ablation study experiment results on the three-target MSTAR data are summarized in Table 8. In Table 8, the addition denotes the element-wise addition of the fusion operation and the concatenation denotes the concatenation fusion operation, which are usually adopted as a fusion module in multi-stream network architectures [66,67]. From rows 1 and 2 in Table 8, it can be observed that recognition accuracy using only the HRRP data through the VAE model was 0.8813 for overall accuracy and 0.8399 for average accuracy. Moreover, the recognition accuracy using only the SAR images through LightNet was 0.9487 for overall accuracy and 0.9612 for average accuracy. The VAE model and the LightNet can extract different features from different domains, and both of them gain good recognition performance. Nevertheless, when comparing rows 1 and 2 with rows 3, 4, 5 and 6, it can be observed that the fusion of the latent feature of the HRRP data obtained from the VAE and the structure features of the SAR images obtained from LightNet are beneficial for reflecting target information more comprehensively and sufficiently to achieve better recognition performance. Furthermore, as shown in rows 3, 4, 5 and 6, on the basis of fusing VAE and LightNet, the performance improvements brought by the different fusion modules were different. The decision-level fusion module had 0.9278 overall accuracy and 0.9357 average accuracy, which were lower than the accuracy only using LightNet. In fact, simple decision-level fusion can indeed get robust performance but finds it difficult to obtain the best performance. The utilization of the element-wise addition module had an 0.9568 overall accuracy and 0.9664 average accuracy; the concatenation module had an 0.9648 overall accuracy and an 0.9715 average accuracy, and the proposed fusion module produced markedly superior recognition accuracy of 0.9780 for overall accuracy and 0.9807 average accuracy. From the comparison we can see that the proposed fusion module achieved the best fusion performance.

Table 8. Ablation Study.

VAE Stream	LightNet Stream	Fusion Module				Overall Accuracy	Average Accuracy
		Decision Level Fusion	Addition	Concatenation	Proposed Fusion Module		
√	×	×	×	×	×	0.8813	0.8399
×	√	×	×	×	×	0.9487	0.9612
√	√	√	×	×	×	0.9278	0.9357
√	√	×	√	×	×	0.9568	0.9664
√	√	×	×	√	×	0.9648	0.9715
√	√	×	×	×	√	**0.9780**	**0.9807**

4.5.2. Feature Analysis

The quantitative performance analysis has been evaluated through comparisons to existing methods and detailed ablation studies to reveal the effectiveness of the proposed method. In this sub-section, we adopt t-SNE [68] to visualize the fusion feature learned by the proposed method, the features learned through the VAE model and the LightNet, as well as the amplitude feature of the original SAR images in Figure 9 on the three-target data. From Figure 9, it can be observed that the features learned by the proposed fusion network show a better feature distribution, in which each class gathers more closely and the margin between them is much more distinct when compared with other features.



(**a**) (**b**) (**c**) (**d**)

Figure 9. T-SNE visualization of the learned features for (**a**) the original amplitude feature, (**b**) the VAE model, (**c**) LightNet, and (**d**) the proposed fusion network.

4.5.3. FLOPs Analysis

In Table 9, we give the floating point of operations (FLOPs) for the VAE branch, the LightNet branch, the proposed network and the VGG network for comparison.

Table 9. FLOPs.

	VAE Branch	LightNet Branch	Proposed Network	VGG Network	ResNet-18	ResNet-34	DenseNet
FLOPs	3.4×10^5	2.3×10^7	2.33525×10^7	5.14×10^9	1.9×10^9	3.6×10^9	5.7×10^9

By analyzing the calculation principle of the convolutional layer, we get that the computational complexity of one convolutional layer is $C_{in,cl}C_{out,cl}{K_{cl}}^2{M_{out,cl}}^2$, where $C_{in,cl}$ and $C_{out,cl}$ are the number of channels in the input and output feature map of the convolutional layer, K_{cl} is the size of the convolution kernel, and $M_{out,cl}$ is the size of the output feature map. For one fully connected layer, the computational complexity is $N_{in.fl}N_{out,fl}$, where $N_{in.fl}$ is the number of input nodes of the fully connected layer and $N_{out,fl}$ is the number of output nodes. Therefore, according to the architecture and details of the LightNet branch shown in Table 1, we obtain the FLOPs for the LightNet branch as 2.3×10^7 by substituting the relevant parameters into the formula of computational complexity. Similarly, according to the introduction of the VAE and the detail of

its architecture presented in Section II-B, we can obtain the FLOPs for the VAE branch as 3.4×10^5. In the proposed network, besides the LightNet branch and the VAE branch, there is a fusion module with 1.25×10^4 FLOPs. Therefore, the total FLOPs of the proposed network are 2.33525×10^7. By substituting the relevant parameters in the VGG network, ResNet-18, ResNet-34 and DenseNet, we can obtain that the FLOPs for VGG were 5.14×10^9, 1.9×10^9, 3.6×10^9 and 5.7×10^9, respectively.

It can be seen from Table 9 that although the VGG network, ResNet-18, ResNet-34 and DenseNet have deeper architectures and require more FLOPs, the recognition performance of these methods on all datasets was lower than that of the proposed method.

4.6. Experiments on Civilian Vehicle Dataset

The civilian vehicles dataset was provided by the U.S. Air Force Research Laboratory. The sensor collecting the civilian vehicles data is a high-resolution Circular SAR and the wave band is X-band. The civilian vehicles data includes ten different civilian vehicle targets, i.e., Toyota Camry, Honda Civic 4dr, 1993 Jeep, 1999 Jeep, Nissan Maxima, Mazda MPV, Mitsubishi, Nissan Sentra, Toyota Avalon and Toyota Tacoma. The aspect angles cover from 0° to 360°, and the depression angles of the samples for each target category is 30°. The HH channel as used for training and the VV channel was used for testing. The number of training and test samples in each category were 360. Importantly, the provided data were high-range resolution radar echoes. For the proposed method, the HRRP data and real-valued SAR images were obtained according to the procedure shown in Figure 1.

We compared the performance of the proposed method with some SAR target recognition methods, including directly applying linear SVM to the original SAR images, PCA followed by linear SVM, the template matching method, DL-JDSR, AE, DAE, the VGG network, A-ConvNet, MBGH+CNN with EFF, ResNet-18, ResNet-34 and DenseNet in Figure 10 and Table 10. Here, due to the number of test samples for each category being the same, the overall accuracy and the average accuracy are the same, too, as formulated in Equations (19) and (20). Thereby, only the total accuracy is listed in Table 10. As shown in Figure 10 and Table 10, our proposed method outperforms all the other compared methods. Especially for the 1993 Jeep, 1999 Jeep and Toyota Avalon, the proposed method yielded the highest accuracy. For the other categories, the accuracy of our method was not the highest, but it was also among the best. And in terms of total accuracy, we can see that the proposed method was at least 2.1% higher than the other compared methods.

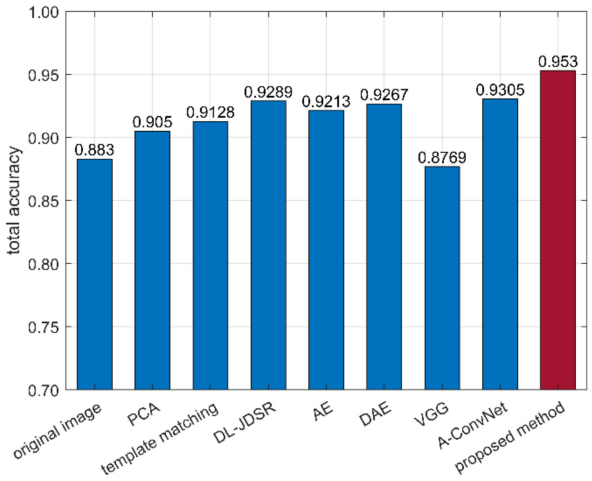


Figure 10. Ten-target accuracy on civilian vehicle data obtained by different SAR target recognition methods.

Table 10. Detailed Accuracies of Different Types on Civilian Vehicle Data via Some SAR Recognition Methods.

	Toyota Camery	Honda Civic 4dr	1993 Jeep	1999 Jeep	Nissan Maxima	Mazda MPV	Mitsu-Bishi	Nissan Sentra	Toyota Avalon	Toyota Tacoma	Total Accuracy
Proposed method	0.8694	0.9472	**1**	**0.9306**	0.9639	0.9528	0.9111	0.9889	**1**	0.9667	**0.9530**
Original image	0.9666	0.9306	0.9861	0.6528	0.7833	**1**	0.9	0.6111	**1**	**1**	0.8830
PCA	0.9444	0.9389	0.9556	0.6944	0.8278	**1**	0.9583	0.7306	**1**	**1**	0.9050
Template matching	0.9083	0.9194	0.9389	0.8722	0.9167	0.9444	0.8639	0.8333	0.9444	0.9861	0.9128
DL-JDSR	0.8833	0.9806	0.9750	0.9111	0.9639	0.8222	0.9583	0.9444	0.85	**1**	0.9289
AE	**0.9944**	0.9639	0.9389	0.8722	0.8778	0.9694	**0.9639**	0.6333	**1**	**1**	0.9213
DAE	0.9889	0.9722	0.9917	0.85	0.8833	0.9722	0.9278	0.6833	0.9972	**1**	0.9267
VGGNet	0.8278	0.7694	0.9611	0.7	0.9361	0.9139	0.8417	0.9194	0.9750	0.9250	0.8769
A-ConvNets	0.8694	0.9306	0.9972	0.8444	**0.9917**	0.9528	0.7306	**0.9972**	**1**	0.9917	0.9305
ResNet-18	0.9452	0.9345	0.9764	0.8857	0.9248	0.9934	0.7756	0.7911	**1**	0.9847	0.9211
ResNet-34	0.9437	0.9647	0.9713	0.6793	0.9537	0.9769	0.8985	0.8865	0.9691	**1**	0.9247
DenseNet	0.9608	0.9762	0.9842	0.9136	0.9157	0.9845	0.8223	0.7567	**1**	**1**	0.9314
MBGH+CNN with EFF	0.8757	**0.9827**	0.9801	0.8348	0.9214	0.9187	0.8467	0.8709	0.9422	0.9341	0.9017

5. Conclusions

In this paper, considering that both SAR images and the corresponding HRRP data, in which the information contained is not exactly the same, can be simultaneously obtained in the procedure of SAR imaging, we formulated a novel end-to-end two stream fusion network framework to fuse the characteristics obtained from modeling HRRP data and SAR images for radar target recognition. The proposed fusion network contains two separated streams in the feature extraction stage, one of which takes advantage of a VAE network to acquire latent probabilistic distribution from the HRRP data and the other using LightNet to extract 2D visual structure information based on the SAR images. The proposed fusion module was utilized to integrate the above-mentioned two types of different characteristics to reflect target information more comprehensively and sufficiently, and it could also merge the two streams into a unified framework with end-to-end joint training. The experimental results based on the MSTAR dataset and the civilian vehicle dataset show that the proposed two-stream fusion method has greater performance advantages than some conventional target recognition methods and other deep learning-based target recognition methods, showing the superiority of the proposed method.

Although the proposed target recognition method offers a significant improvement in performance, there is also a limit in speed. Since the proposed method contains two branches, the running time of the proposed method on one test sample is a little higher than that of a single branch. In the future, we will further explore the increase in speed draw support through parallel computing and algorithm optimization.

Author Contributions: Conceptualization, L.D.; methodology, L.D.; software, L.L. and Y.G.; validation, L.L., K.R., J.C., L.D. and Y.G.; investigation, Y.W.; resources, L.D.; writing—original draft preparation, L.L.; writing—review and editing, Y.G. and L.D.; visualization, L.L.; supervision, L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science Foundation of China under Grant 61771362 and in part by the 111 Project.

Data Availability Statement: Not applicable.

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

References

1. Chen, S.; Wang, H. SAR target recognition based on deep learning. In Proceedings of the International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, 19–21 October 2015.
2. Cui, Z.; Cao, Z.; Yang, J.; Ren, H. Hierarchical Recognition System for Target Recognition from Sparse Representations. *Math. Probl. Eng.* **2015**, *2015 Pt 17*, 6. [CrossRef]
3. Deng, S.; Du, L.; Li, C.; Ding, J.; Liu, H. SAR automatic target recognition based on euclidean distance restricted autoencoder. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3323–3333. [CrossRef]
4. Housseini, A.E.; Toumi, A.; Khenchaf, A. Deep Learning for Target recognition from SAR images. In Proceedings of the 2017 Seminar on Detection Systems Architectures and Technologies (DAT), Algiers, Algeria, 20–22 February 2017.
5. Yan, H.; Zhang, Z.; Gang, X.; Yu, W. Radar HRRP recognition based on sparse denoising autoencoder and multi-layer perceptron deep model. In Proceedings of the 2016 Fourth International Conference on Ubiquitous Positioning, Indoor Navigation and Location Based Services (UPINLBS), Shanghai, China, 2–4 November 2016.
6. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
7. Du, C.; Chen, B.; Xu, B.; Guo, D.; Liu, H. Factorized Discriminative Conditional Variational Auto-encoder for Radar HRRP Target Recognition. *Signal Process.* **2019**, *158*, 176–189. [CrossRef]
8. Du, L.; Liu, H.; Bao, Z.; Zhang, J. Radar automatic target recognition using complex high-resolution range profiles. *IET Radar Sonar Navig.* **2007**, *1*, 18–26. [CrossRef]
9. Du, L. Noise Robust Radar HRRP Target Recognition Based on Multitask Factor Analysis With Small Training Data Size. *IEEE Trans. Signal Process.* **2012**, *60*, 3546–3559.
10. Xing, M. Properties of high-resolution range profiles. *Opt. Eng.* **2002**, *41*, 493–504. [CrossRef]
11. Zhang, X.Z.; Huang, P.K. Multi-aspect SAR target recognition based on features of sequential complex HRRP using CICA. *Syst. Eng. Electron.* **2012**, *34*, 263–269.
12. Masahiko, N.; Liao, X.J.; Carin, L. Target identification from multi-aspect high range-resolution radar signatures using a hidden Markov model. *IEICE Trans. Electron.* **2004**, *87*, 1706–1714.
13. Tan, X.; Li, J. Rang-Doppler imaging via forward- backward sparse Bayesian learning. *IEEE Trans. Signal Process.* **2010**, *58*, 2421–2425. [CrossRef]
14. Zhao, F.; Liu, Y.; Huo, K.; Zhang, S.; Zhang, Z. Radar HRRP Target Recognition Based on Stacked Autoencoder and Extreme Learning Machine. *Sensors* **2018**, *18*, 173. [CrossRef] [PubMed]
15. Feng, B.; Chen, B.; Liu, H. Radar HRRP target recognition with deep networks. *Pattern Recognit.* **2017**, *61*, 379–393. [CrossRef]
16. Pan, M.; Liu, A.L.; Yu, Y.Z.; Wang, P.; Li, J.; Liu, Y.; Lv, S.S.; Zhu, H. Radar HRRP target recognition model based on a stacked CNN-Bi-RNN with attention mechanism. *IEEE Trans. Geosci. Remote Sens.* **2021**, *61*, 1–14, online published. [CrossRef]
17. Chen, W.C.; Chen, B.; Peng, X.J.; Liu, J.; Yang, Y.; Zhang, H.; Liu, H. Tensor RNN with Bayesian nonparametric mixture for radar HRRP modeling and target recognition. *IEEE Trans. Signal Process.* **2021**, *69*, 1995–2009. [CrossRef]
18. Peng, X.; Gao, X.Z.; Zhang, Y.F. An adaptive feature learning model for sequential radar high resolution range profile recognition. *Sensors* **2017**, *17*, 1675. [CrossRef] [PubMed]
19. Jacobs, S.P. *Automatic Target Recognition Using High-Resolution Radar Range-Profiles*; ProQuest Dissertations Publishing: Morrisville, NC, USA, 1997.
20. Webb, A.R. Gamma mixture models for target recognition. *Pattern Recognit.* **2000**, *33*, 2045–2054. [CrossRef]
21. Copsey, K.; Webb, A. Bayesian gamma mixture model approach to radar target recognition. *IEEE Trans. Aerosp. Electron. Syst.* **2003**, *39*, 1201–1217. [CrossRef]
22. Du, L.; Liu, H.; Zheng, B.; Zhang, J. A two-distribution compounded statistical model for Radar HRRP target recognition. *IEEE Trans. Signal Process.* **2006**, *54*, 2226–2238.
23. Du, L.; Liu, H.; Bao, Z. Radar HRRP Statistical Recognition: Parametric Model and Model Selection. *IEEE Trans. Signal Process.* **2008**, *56*, 1931–1944. [CrossRef]
24. Du, L.; Wang, P.; Zhang, L.; He, H.; Liu, H. Robust statistical recognition and reconstruction scheme based on hierarchical Bayesian learning of HRR radar target signal. *Expert Syst. Appl.* **2015**, *42*, 5860–5873. [CrossRef]
25. Park, S.C.; Park, M.K.; Kang, M.G. Super-Resolution Image Reconstruction: A Technical Overview. *IEEE Signal Process. Mag.* **2003**, *20*, 21–36. [CrossRef]
26. Wang, P.; Shi, L.; Lan, D.; Liu, H.; Xu, L.; Bao, Z. Radar HRRP Statistical Recognition With Local Factor Analysis by Automatic Bayesian Ying-Yang Harmony Learning. *Front. Electr. Electron. Eng. China* **2011**, *6*, 300–317. [CrossRef]
27. Chen, J.; Du, L.; He, H.; Guo, Y. Convolutional factor analysis model with application to radar automatic target recognition. *Pattern Recognit.* **2019**, *87*, 140–156. [CrossRef]
28. Pan, M.; Du, L.; Wang, P.; Liu, H.; Bao, Z. Noise-Robust Modification Method for Gaussian-Based Models With Application to Radar HRRP Recognition. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 55–62. [CrossRef]
29. Chen, H.; Guo, Z.Y.; Duan, H.B.; Ban, D. A genetic programming-driven data fitting method. *IEEE Access* **2020**, *8*, 111448–111458. [CrossRef]
30. Rezende, D.J.; Mohamed, S.; Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
31. Doersch, C. Tutorial on Variational Autoencoders. *arXiv* **2016**, arXiv:1606.05908.

32. Ying, Z.; Bo, C.; Hao, Z.; Wang, Z. Robust Variational Auto-Encoder for Radar HRRP Target Recognition. In Proceedings of the International Conference on Intelligent Science & Big Data Engineering, Dalian, China, 22–23 September 2017.
33. Chen, J.; Du, L.; Liao, L. Class Factorized Variational Auto-encoder for Radar HRRP Target Recognition. In Proceedings of the 2020 IEEE Radar Conference (RadarConf20), Florence, Italy, 21–25 September 2020.
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
35. Min, R.; Lan, H.; Cao, Z.; Cui, Z. A Gradually Distilled CNN for SAR Target Recognition. *IEEE Access* **2019**, *7*, 42190–42200. [CrossRef]
36. Huang, X.; Yang, Q.; Qiao, H. Lightweight two-stream convolutional neural network for SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 667–671. [CrossRef]
37. Cho, J.; Chan, G. Multiple feature aggregation using convolutional neural networks for SAR image-based automatic target recognition. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1882–1886. [CrossRef]
38. Ruser, H.; Leon, F.P. Information fusion—An overview. *Tech. Mess.* **2006**, *74*, 93–102. [CrossRef]
39. Jiang, L.; Yan, L.; Xia, Y.; Guo, Q.; Fu, M.; Lu, K. Asynchronous multirate multisensor data fusion over unreliable measurements with correlated noise. *IEEE Trans. Aerosp. Electron. Syst.* **2017**, *53*, 2427–2437. [CrossRef]
40. Rasti, B.; Ghamisi, P.; Plaza, J.; Plaza, A. Fusion of hyperspectral and LiDAR data using sparse and low-rank component analysis. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 6354–6365. [CrossRef]
41. Bassford, M.; Painter, B. Intelligent bio-environments: Exploring fuzzy logic approaches to the honeybee crisis. In Proceedings of the 2016 12th International Conference on Intelligent Environments (IE), London, UK, 14–16 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 202–205.
42. Mehra, A.; Jain, N.; Srivastava, H.S. A novel approach to use semantic segmentation based deep learning networks to classify multi-temporal SAR data. *Geocarto Int.* **2020**, 1–16. [CrossRef]
43. Pei, J.; Huang, Y.; Huo, W.; Zhang, Y.; Yang, J.; Yeo, T.S. SAR automatic target recognition based on multiview deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2196–2210. [CrossRef]
44. Choi, I.O.; Jung, J.H.; Kim, S.H.; Kim, K.T.; Park, S.H. Classification of targets improved by fusion of range profile and the inverse synthetic aperture radar image. *Prog. Electromagn. Res.* **2014**, *144*, 23–31. [CrossRef]
45. Wang, L.X.; Weng, L.G.; Xia, M.; Liu, J.; Lin, H. Multi-resolution supervision network with an adaptive weighted loss for desert segmentation. *Remote Sens.* **2021**, *13*, 1–18.
46. Shang, R.H.; Zhang, J.Y.; Jiao, L.C.; Li, Y.; Marturi, N.; Stolkin, R. Multi-scale Adaptive feature fusion network for segmentation in remote sensing images. *Remote Sens.* **2020**, *12*, 872. [CrossRef]
47. Chen, J.; He, F.; Zhang, Y.; Sun, G.; Deng, M. SPMF-net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion. *Remote Sens.* **2020**, *12*, 1049. [CrossRef]
48. Liao, X.; Runkle, P.; Carin, L. Identification of ground targets from sequential high-range-resolution radar signatures. *IEEE Trans. Aerosp. Electron. Syst.* **2002**, *38*, 1230–1242. [CrossRef]
49. Zhang, X.; Liu, Z.; Liu, S.; Li, G. Time-Frequency Feature Extraction of HRRP Using AGR and NMF for SAR ATR. *J. Electr. Comput. Eng.* **2015**, *2015*, 340–349. [CrossRef]
50. Chen, B.; Liu, H.; Bao, Z. Analysis of three kinds of classification based on different absolute alignment methods. *Mod. Radar* **2006**, *28*, 58–62.
51. Lan, D.; Liu, H.; Zheng, B.; Xing, M. Radar HRRP Target Recognition Based on Higher Order Spectra. *IEEE Trans. Signal Process.* **2005**, *53*, 2359–2368. [CrossRef]
52. Beal, M. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, University College London, London, UK, 2003.
53. Nielsen, F.B. Variational Approach to Factor Analysis and Related Models. Master's Thesis, Informatics and Mathematical Modelling, Technical University of Denmark, Copenhagen, Denmark, 2004.
54. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015.
55. Gulcehre, C.; Cho, K.; Pascanu, R.; Bengio, Y. Learned-Norm Pooling for Deep Feedforward and Recurrent Neural Networks. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Nancy, France, 15–19 September 2014.
56. The Sensor Data Management System. Available online: https://www.sdms.afrl.af.mil/index.php?collection=mstar (accessed on 10 September 2015).
57. Sun, Y.; Du, L.; Wang, Y.; Wang, Y.; Hu, J. SAR automatic target recognition based on dictionary learning and joint dynamic sparse representation. *IEEE Geosci. Remote Sens. Lett.* **2017**, *13*, 1777–1781. [CrossRef]
58. Dong, G.; Liu, H.; Kuang, G.; Chanussot, J. Target recognition in SAR images via sparse representation in the frequency domain. *Pattern Recognit.* **2019**, *96*, 106972. [CrossRef]
59. Dong, G.; Kuang, G. Target recognition in SAR images via classification on Riemannian manifolds. *IEEE Geosci. Remote Sens. Lett.* **2014**, *12*, 199–203. [CrossRef]
60. Chen, S.; Wang, H.; Xu, F.; Jin, Y.Q. Target Classification Using the Deep Convolutional Networks for SAR images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1–12. [CrossRef]

61. Theagarajan, R.; Bhanu, B.; Erpek, T.; Hue, Y.K.; Schwieterman, R.; Davaslioglu, K.; Shi, Y.; Sagduyu, Y.E. Integrating deep learning-based data driven and model-based approaches for inverse synthetic aperture radar target recognition. *Opt. Eng.* **2020**, *59*, 051407. [CrossRef]
62. Guo, J.; Wang, L.; Zhu, D.; Hu, C. Compact convolutional autoencoder for SAR target recognition. *IET Radar Sonar Navig.* **2020**, *14*, 967–972. [CrossRef]
63. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
64. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
65. Yu, M.; Dong, G.; Fan, H.; Kuang, G. SAR Target Recognition via Local Sparse Representation of Multi-Manifold Regularized Low-Rank Approximation. *Remote Sens.* **2018**, *10*, 211. [CrossRef]
66. Mou, L.; Schmitt, M.; Wang, Y.; Zhu, X.X. A CNN for the identification of corresponding patches in SAR and optical imagery of urban scenes. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates, 6–8 March 2017.
67. Hu, J.; Mou, L.; Schmitt, A.; Zhu, X.X. FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, United Arab Emirates, 6–8 March 2017.
68. Laurens, V.D.M.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.