Optik - International Journal for Light and Electron Optics 185 (2019) 1104–1114

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.com/locate/ijleo

Original research article

Concealed object segmentation in terahertz imaging via adversarial learning

Dong Liang[a,b,*], Jiaxing Pan[a,b], Yang Yu[c,d], Huiyu Zhou[e]

[a] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[b] Collaborative Innovation Center of Novel Software Technology and Industrialization, China
[c] Microsystem and Terahertz Research Center, China Academy of Engineering Physics, China
[d] Institue of Electronic Engineering, China Academy of Engineering Physics, China
[e] Department of Informatics, University of Leicester, United Kingdom

ARTICLE INFO

Keywords:
Terahertz
Generative adversarial nets
Object segmentation

ABSTRACT

Terahertz imaging (frequency between 0.1 to 10 THz) is a modern technique for public security check. Due to poor imaging quality, traditional machine vision methods often fail to detect concealed weapons in Terahertz samples, while modern instance segmentation approaches have complex multiple-stage concatenation and often hunger for massive and accurate training data. In this work, we realize a novel Conditional Generative Adversarial Nets (CGANs), named as Mask-CGANs to segment weapons in such a challenging imaging quality. The Mask-Generator network employs a "selected-connection U-Net" to restrain false alarms and speed up training convergence. The loss function takes reconstruction errors and sparse priors into consideration to preserve precise segmentation. Such a learning architecture works well with a small training dataset. Experiments show that the proposed model outperforms CGANs (more than 16–32% in Recall, Precision and Accuracy) and Mask-RCNN (more than 3–6%). Moreover, its testing speed (69.7 *FPS*) is fast enough to be implemented in a real-time security check system, which is 44 times faster than Mask-RCNN. In the experiments for mammographic mass segmentation on INBreast dataset, the Dice index of the proposed method is 91.29, surpasses the-state-of-the-art medical issue segmentation methods. The full implementation (based on TensorFlow) is available at https://github.com/JXPanzz/THz).

1. Introduction

Detecting concealed objects underneath clothing is a critical task in public security check, while the traditional manual check is often criticized with inefficiency, invasion of privacy, and high rates of missed detection. Terahertz imaging technology [1–3], provides a non-contact and non-destructive way to acquire samples of objects concealed underneath clothing with no harmful to health. However, due to the inherent physical properties of Terahertz imaging, samples have low contrast and signal to noise ratio. Fig. 1 shows our Terahertz imaging samples, where the concealed weapons are inconspicuous in intensity value, yet the boundary of objects are partially indistinguishable from the human body.

A typical use of modern deep convolutional networks is on classification tasks, where the output to an image is a single class label. However, in many visual tasks, especially in a security check system, the desired output should include localization, i.e., a class label

* Corresponding author at: College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.
E-mail address: liangdong@nuaa.edu.cn (D. Liang).

https://doi.org/10.1016/j.ijleo.2019.04.034
Received 25 September 2018; Received in revised form 13 March 2019; Accepted 3 April 2019

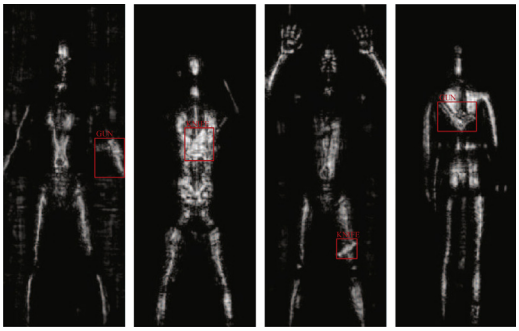

Fig. 1. Images in our dataset. These four samples show weapons concealed in different positions on four subjects from fore and back views (weapons are marked by red bounding boxes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is supposed to be assigned to each pixel. Traditional object recognition often follows the detection-first regime. Unfortunately, the imaging quality mentioned above leads to poor detection. Our idea is to first accurately segment concealed objects from subjects' body without any prior supervised label. Following this way, actually, detection is just the post-processing following segmentation – When the segmentor provides a none-zero blob, then provide a bounding-box to highlight the region. Obviously, this is a instance-segmentation regime. Modern instance segmentation approaches based on deep convolutional neural networks [4–13] have great potential in this task. But segmentation and detection often belong to two relatively independent units, which makes the modeling structure rather complex resulting in a slow processing speed and less robustness. Moreover, modern deep convolutional networks often hunger for massive training data. Since the imaging highly depends on specific equipment, thousands of training images for each class are beyond reach in this task.

Generative adversarial nets (GANs) [14] were introduced for training generative models in order to sidestep the difficulty of approximating many intractable probabilistic computations when just relying on a small dataset. It can produce state of the art log-likelihood estimates and realistic samples. A conditional GANs (or CGANs) model [15] by giving the model additional information, is possible to direct the data generation process. Such conditioning could be based on class labels, on some part of data for inpainting, or even on data from different modalities. CGANs have been applied in many image generation tasks, such as image forgery and super resolution.

In this work, we realize this particular instance segmentation task as an image-to-image transformation process that translates input samples into masks containing interested objects while dropping unrelated ones. We propose Mask-CGANs shown in Fig. 2. To meet the specific needs, the loss function of Mask-CGANs takes reconstruction errors and sparse prior knowledge into consideration. In the Mask-Generator net, a "selected-connection U-Net" is designed to preserve objects and discard useless image details which makes a trade-off between Recall and False Alarm rates. Experiments show that the proposed Mask-CGANs outperform CGANs and the state of the art instance segmentation approach Mask-RCNN in both robustness and processing speed.

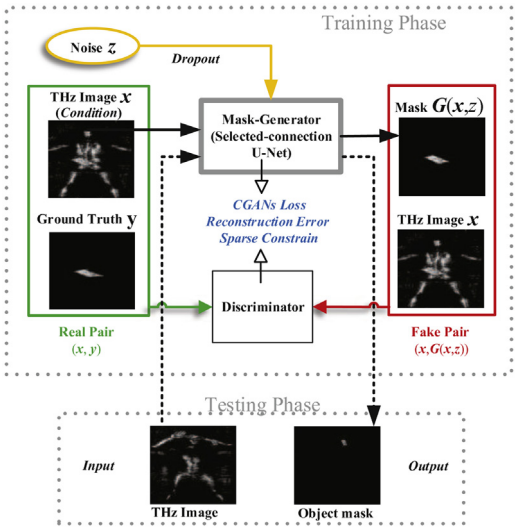

Fig. 2. Mask-CGANs framework for segmenting weapons in Terahertz samples. Image x is added into Mask-Generator as a condition to achieve CGANs. Segmentation masks are generated by Mask-Generator under the supervision of Discriminator and loss function during training phase. Masks are output from the trained Mask-Generator during testing phase.

The contributions of this work are:

- Introducing adversarial learning to deal with a small training dataset.
- Designing a "selected-connection U-Net" for the Mask-Generator to make a balance between Recall and False Alarm rates.
- Introducing reconstruction errors and sparse priors into consideration for the loss function of the proposed Mask-CGANs, to provide precise segmentation.
- It is appropriate to be implemented in a real-time Terahertz security inspection system.

The outline of this paper is organized as follows: the related work will be introduced in Section 2, the proposed Mask-CGANs used in this task will be introduced in detail in Section 3, Dataset preparation and augmentation, experiment set-up and results, and the discussion about results will be presented in Section 4, conclusion and future work will be provided in Section 5.

2. Related work

As the samples got from Terahertz waves are under poor quality, finding the prohibited items from the present low quality electromagnetic imaging data becomes particularly difficult. Early work focused on statistical modeling. Shen [16] proposed multilevel thresholds segmentation by using a Gaussian Mixture Model (GMM) to model the probability density function of radiometric temperature and segment objects by tacking the evolution of boundaries during threshold changing, and the anisotropic diffusion algorithm is applied as a pre-processing to remove noise. Lee [17] builds GMM for the samples and then human bodies and objects are divided by Bayesian boundaries in two Expectation Maximization (EM) processing respectively. Yeom et al [18] applied a multilevel segmentation process based on GMM and vector quantization used before EM for real-time detection. These statistical modelling methods are applicable to segment object in the weak supervised way which lack the capability in accurate detection without any label, and the segmentation performance is poor under bad imaging quality.

Modern general instance segmentation approaches based on deep convolutional networks have great potential in this task. This kind of approaches could be roughly divided into two families. One family relies on the R-CNN proposals, which is a bottom-up pipeline that the segmentation results are based on the proposals and then labelled by a classifier [4–9]. The other family relies on semantic segmentation results [10–13] where instance segmentation following semantic segmentation by classifying pixels into different instances. A state-of-the-art method Mask-RCNN [19], built upon object detectors [20,21], also depends on the proposals but features are shared by classes, box predictors and mask generators, then all results are collected in parallel. However, all of those algorithms contain complex multiple-stage cascading which is slow and less robust. In addition, the deep convolutional networks based framework is typically requiring big training datasets to guarantee their generalization for new data while our dataset is rather small (with 1440 samples in 2 classes for both training and testing).

GANs provides a generative model which can achieve image-to-image mapping. A GANs [14] aims to fit a mapping function f from noise z input to its output sample g, $f: z \rightarrow g$, who does not need the prior knowledge about a latent data distribution and captures it just through an adversarial process. A typical GANs model contains two parts. One is Generator (G) which tries to generate fake samples but more and more realistic then tries to fool the other one called Discriminator (D) to make mistakes. And the assignment for D is to distinguish whether the samples received are fake generated by G or real ones sampled from real data. Then the ultimate outcome is that G can generate samples where D cannot determinate if they are fake or not. A mount of new training data produced during this adversarial process can release pressure on small datasets. Outwardly, it is fully under an unsupervised manner, so that original GANs can be seen just as a noise mixer to generate samples with a latent noise distribution. Additional information is added into the unsupervised model to direct the generating process, for example, CGANS aims at generating images under certain constraints [15,22–26], which could be seen as a supervised label for each training sample. We argue that object segmentation can be regarded as a special sample generation task. The difference is that the generated sample is the mask of the object. Thus, the inner structure of GANs should be designed to meet this particular aim.

3. Model structure of Mask-CGANs

As a supervised segmentation task, the location of the object should be preserved. As it is shown in Fig. 2, during the training phase, Mask-CGANs learns the mapping function f from a Terahertz sample x to its ground truth y. The training is supervised by a Terahertz sample x as condition along with its ground truth y and a noise z produced by the drop-out process in the Mask-Generator model to output a fake sample $G(x, z)$, correspondingly. While the inputs Discriminator receiving include two parts: one is the real pair: a Terahertz sample with its ground truth, the other contains a real Terahertz sample and a fake sample produced by the Mask-Generator. In the testing phase, given a condition sample, the mask sample is generated deterministic by the trained Mask-Generator. For the Discriminator of GANs, we follow the Discriminator net model used in [15], a convolutional PatchGAN classifier, which only penalizes structure at the scale of image patches.

3.1. Selected-connection U-Net for Mask-Generator

For the designing of the Mask-Generator, ablation studies and Occams Razor are guiders for our study. We provide the discussion and explanation why we design a Selected-connection U-net structure.

For the Mask-Generator net, our model and another two candidates are illustrated in Fig. 3(a). As a generator, traditional

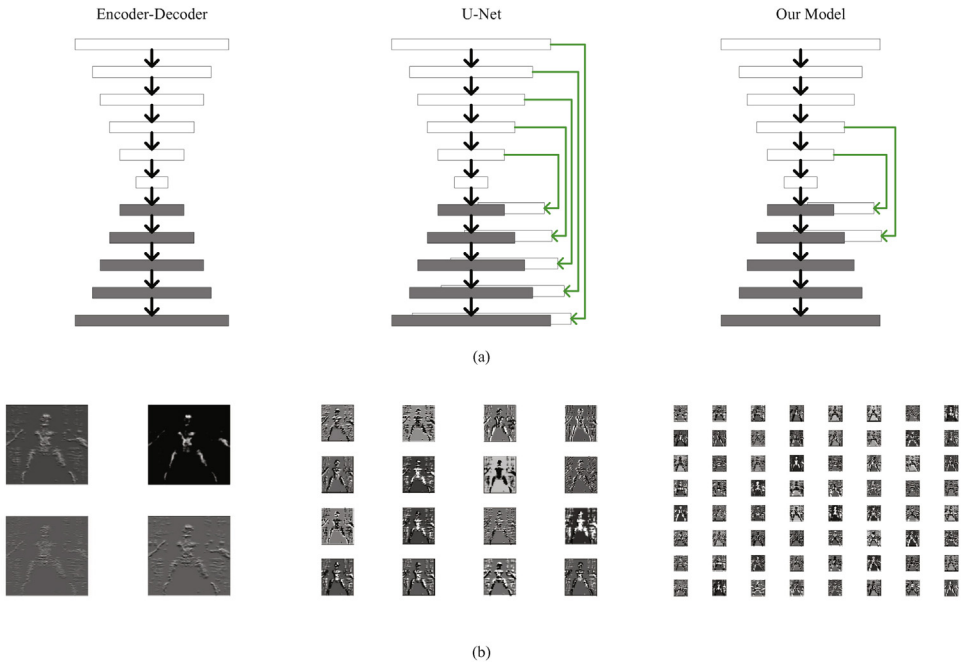

Fig. 3. (a) Mask-Generator "selected-connection U-Net" model and two other candidates. The white blocks mean the down-sampling layers' features and the up-sampling layers' are in gray color. The connection lines (in green) copy features from the beginning layers to terminal layers. (b) We sample some feature maps produced by the first three encoder layers. Rich low-level features like edges are shown in those feature map which will cause a higher False Alarm rate when concatenated to the layers nearby the output layers. So our model drops this feature from "U-net". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

"Encoder-Decoder" [27] is the basic feature mapping model. It represents the mapping function from a Terahertz sample x to its segmentation mask $G(x, z)$. Image x passes through several down-sampling layers until a bottleneck layer and then up-sampling process applied symmetrically. In another word, all information flows through all layers where low-level features are lost. However, there is a great deal of low-level information shared between the input and output, and it would be desirable to shuttle this information directly across the net. For example, in the case of tiny objects segmentation, the input and output share the location of prominent edges, which are important to help find the objects.

"U-Net" [28] employed in [15] concatenates all the layers before the bottleneck layers with richer fine-grained features to hold more realistic samples. In experiment, we found it provides image details, but, produces many false alarms, which can be seen in Table 1. The original "U-net" essentially propagate information from the down-sampling layers to all corresponding symmetrical up-sampling layers. For a segmentation task, the low-level feature layers contain too much redundant interference. However, we should take care of the balance between recalling hidden objects as many as possible and making mistakes as few as possible. As shown in Fig. 3(b), the feature maps sampled from the first three down-sampling layers who carry too much edge information and background which we want to discard when segmenting the objects of interested. Such low-level features may lead to higher false alarms when they emerge in the up-sampling layers nearby the output one.

The above mentioned is the motivation why we propose a "selected-connection U-Net", to discard the feature channels between low-level feature layers, and to preserve the ones between high-level feature layers. The experimental results indicate this improvement works much better than original U-net and Encoder-Decoder. Along with the loss function (will be introduced in Section 3.3), it makes a good tradeoff between Recall and False Alarm rates.

The Mask-Generator, selected-connection U-Net is shown in Fig. 4. This model is a fully convolutional net with 16 layers, and each

Table 1
Overall evaluation comparisons.

Generator or method	Loss function	Recall	Accuracy	Precision	False Alarm
Encoder-Decoder [27]	$L_{\text{CGANs}} + L_{L_1}$	0.6375	0.5083	0.6296	0.7500
U-Net(CGANs) [15]	L_{CGANs}	0.7500	0.5000	0.6000	1.0000
U-Net	$L_{\text{CGANs}} + L_{L_1}$	0.9125	0.7750	0.7849	0.5000
selected-connection U-Net	$L_{\text{CGANs}} + L_{L_1}$	0.9125	0.8000	0.8111	0.4250
Mask-RCNN [19]	–	0.8563	0.7583	0.7965	0.4375
selected-connection U-Net	$L_{\text{CGANs}} + L_{L_1} + L_S$	**0.9125**	**0.8208**	**0.8343**	**0.3625**

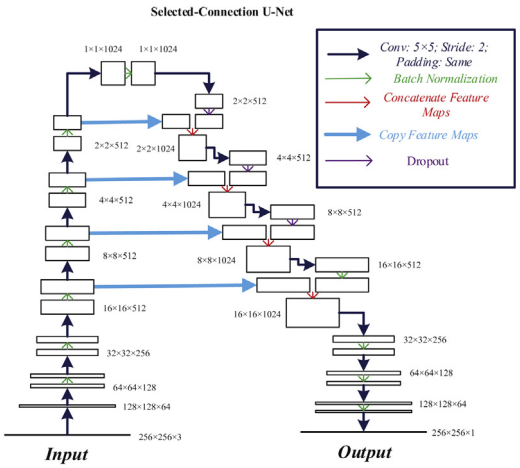

Fig. 4. The model structure of selected-connection U-Net. The model is a fully convolutional net with 16 layers, each layer has a 5*5 convolution kernel with stride equal to 2*2 and padding.

layer has a 5*5 convolution kernel with stride equal to 2*2 and padding. So the image size will be reduced by halve after each process of downsampling which consists 8 convolution layers and Leaky Relu function is taken as activation function. In contrast to the downsampling, the image size is twice as large as the upsampling, but the number of features is halved. When up-sample, selected output feature maps are merged with the corresponding feature maps from the down-sampling network to make well trade-off on richer fine-grained features and lower false alarm rate.

3.2. The role of noise z

Input z is Gaussian noise, a standard operation in GANs, applied to the Mask-Generator during the training process. Without z, the net could still learn a mapping from x to $G(x)$, but would produce deterministic outputs, and therefore fail to match any distribution other than a loss function. For our models, we provide noise only in the form of drop-out, applied on layers of the Mask-Generator. The drop-out noise show minor stochasticity in the output of the Mask-Generator and thereby capture the entropy of the conditional distributions it model. The introducing of Gaussian noise z allows the model to ignore the prior knowledge about the latent data distribution and to captures it just through an adversarial process. New training data is produced during this adversarial process which can release the pressure of overfitting on a small dataset. It makes the trained model more robust in our task.

3.3. Loss function

3.3.1. Basic CGANs loss

To segment concealed objects in a Terahertz sample comes, the basic loss function comes from CGANs [15]. Compared to traditional GANs, these approaches applied certain restrictions. In this work, the condition x (Terahertz sample) is added into each net. The loss function is:

$$L_{\mathrm{CGANs}} = E_{x,y\sim \mathrm{data}(x,y)}\log \mathrm{D}(x, y) + E_{x\sim \mathrm{data}(x),z\sim p(z)}\log(1 - D(x, G(x, z))) \quad (1)$$

In this loss function, both the Mask-Generator and Discriminator have a chance to observe the real input samples x during the training phase. Discriminator maximizes this Expectation term (E) by outputting a high score for the pair (x, y) sampled from distribution $data(x, y)$ while giving the fake pair $(x, G(x, z))$, where x is sampled from real data distribution $data(x)$ and z is produced by $p(z)$, a low score (the score measures the possibility of the sample the Discriminator received to be sampled from real data, and the higher the more realistic). However, the Mask-Generator wants to optimise this function and generate fake samples as realistic as possible to get a higher score for the fake pair. $logD(x, y)$ and $logD(x, G(x, z))$ indicate the logarithm scores for real pair (x, y) and fake pair $(x, G(x, z))$, respectively. In practice, z is a drop-out process in Mask-Generator layers to generate variant outputs. More new samples are generated help to capture latent distributions and enhance the robustness. Condition x is employed to ensure the output mask $G(x, z)$ is in pair with condition x and locate at the corresponding position. Without condition x, the mask would be generated in any arbitrary position.

3.3.2. Reconstruction error

Since the Mask-Generator is an encoder-decoder type generative model and aims to take in noise-perturbed input and reconstruct original input, the loss function take reconstruction error into consideration. The reconstruction error measures the Manhattan distance between the ground truth and the generated one in the pixel level. This part directs the generation process under the shape constrains and enhance the supervision efficiency that restricts the mask output. Also, We use L_1 instead of L_2 as the latter brings

more blurs [15].This term is expressed as:

$$L_{L_1} = \|y - G(x, z)\|_1 \tag{2}$$

3.3.3. Sparse constrain

Here, we also take sparse priors as a constrain term where L_1-norm is applied instead of theoretical L_0-norm or common L_2-norm that avoids solving a NP hard problem in the former norm whilst L_2-norm cannot make most values to zero but small numbers. The sparse prior constrain is measured as:

$$L_S = \|G(x, z)\|_1 \tag{3}$$

The sparse prior helps to keep the generated mask within the bounds and reduce False Alarms thereby. The sparse assumption is made for the targeted object just occupy a small area in the whole sample. Multiple targeted objects do not mean the area will cover most area of the sample. So the sparse constraint also work with multiple targeted objects.

Then the final loss function is reorganized as three main parts:

$$\text{Goal}(G, D) = \arg\min_{G}\max_{D} L_{\text{CGANs}} + \lambda L_{L_1} + \beta L_S \tag{4}$$

Condition x and the reconstruction error provided as supervised terms ensures the generated result to be in pair with the input Terahertz sample. Otherwise, the generated results will be much casual and diversified that is unsuitable for our needs. And precise segmentation performance can be guaranteed by minimizing the distance between mask outputs and its ground truth. What's more, sparse constrain, as prior knowledge, reduces False Alarm rates. The functions of these terms are shown and discussed in Section 4.

4. Experiments

In this section, we apply our Mask-CGANs and other models to segmenting hidden objects in Terahertz samples. Experimental results will be demonstrated and discussed in detail.

4.1. Dataset preparation

The dataset contains 1440 Terahertz samples. Those samples are sampled from 4 subjects (360 for each one including fore and back views) containing weapons like guns, knives or nothing. In experiments we do not distinguish the sample between fore or back view and kinds of weapons, with 2 classes (with weapon or not). We define samples without any hidden weapon as negative while others are positive. Ground truth labels are provided for each sample by manually segmenting the weapon areas by three guys, and using the majority voting to produce the final ground truth. For the positive samples with weapons, segmentation masks are applied as the ground truth, while for the negative samples which do not contain any weapons, their ground truths are the black background.

4.2. Training dataset augmentation

Data augmentation is essential to teach the network the desired invariance and robustness properties, when only few training samples are available. We primarily undertake shift and rotation invariance as well as robustness to deformations and gray value variations, to increase the variants of the trained data during the training phase. Especially random deformations and gray value variations of the training samples seem to be the key concept to train a segmentation network with very few annotated images. We generate smooth deformations using random displacement vectors on a coarse 3 by 3 grid. The displacements are sampled from a Gaussian distribution with the 10 pixels standard deviation. Per-pixel displacements are then computed using bicubic interpolation. Drop-out layers at the end of the contracting path perform further implicit data augmentation.

4.3. Experiment settings

In practice, 2/3 of the samples with their ground truths, selected randomly, are put into the Mask-CGANs model for training. While the remaining 1/3 samples are used for testing. In this experimental evaluation, the validation set and training set are overlapped. Although this way seems to be redundant for training, it is a safe way when the training set is relatively small.

The evaluation methodology is based on “overlapping threshold” [29]. It is strict and rigorous:

(1) A image sample is true positive (TP) if and only if its ground truth blob (connected region containing at least one pixel) is positive, and the overlapping area between the ground truth blob and the generated mask blob is over a predetermined threshold ratio of the total object domain. Otherwise it is regarded as false negative (FN) sample.
(2) The true negative (TN) sample is defined as the ground truth is negative and the output should not have any positive pixels (each pixel value equals to 0). Otherwise it is regarded as false positive (FP) or false alarm sample. Namely, if there is one false positive pixel in the image sample, the sample will be counted as a FP sample.

For quantitative analysis, four evaluation metrics, *Precision*, *Recall*, *Accuracy* and *False Alarm* were utilized,

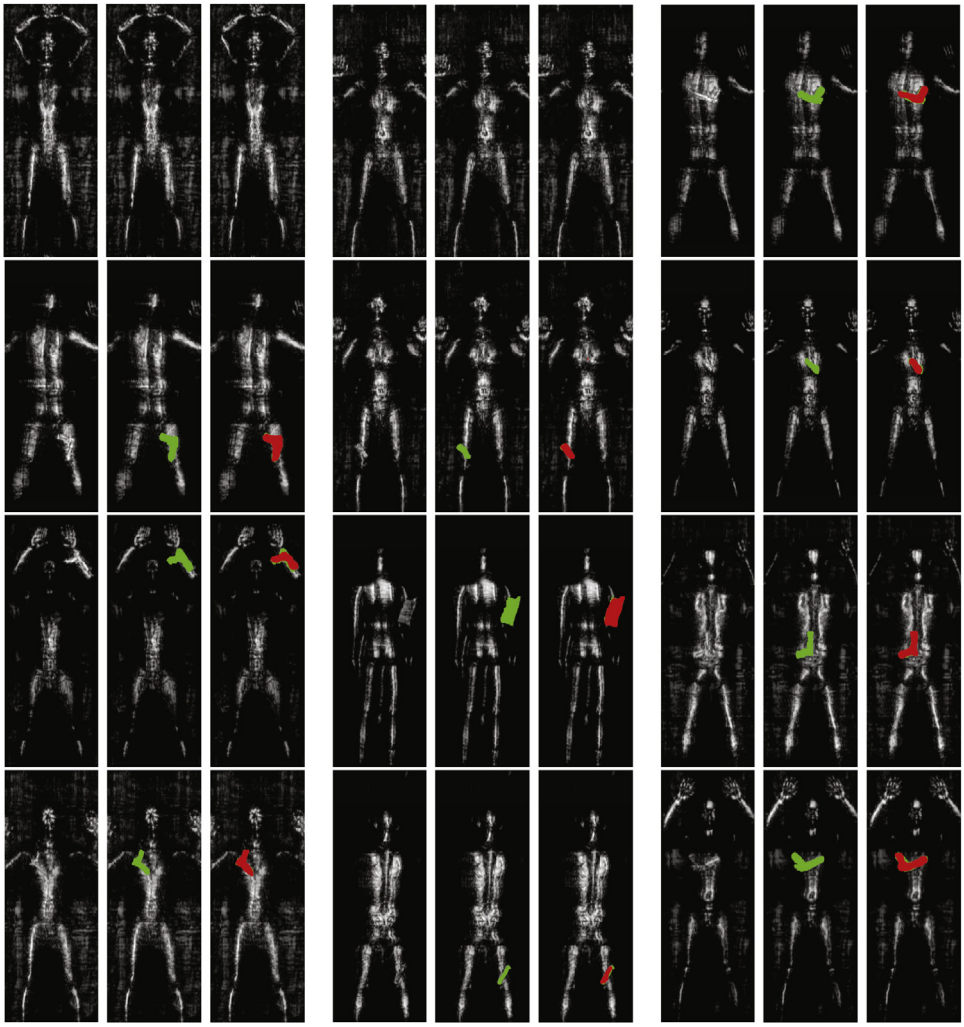

Fig. 5. Experiment results (12 groups). For each sample, it has three columns illustrated the original Terahertz sample (left), sample with ground truth (middle, marked with green area), sample with ground truth and output of Mask-CGANs model (red area) from left to right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{P + N} \tag{7}$$

$$\text{FalseAlarm} = \text{FalsePositiveRate} = \frac{\text{FP}}{N} \tag{8}$$

where P and N represents the number of all positive and negative image samples, respectively. Note that, the False Alarm will not be affected while the threshold changes because if there is one false positive pixel, the image sample will be counted as a FP case.

A compared method in the experiment is Mask-RCNN [19], a state-of-the-art method for instance segmentation. It bases on ResNet-101 [30], and it was pre-trained on COCO dataset. In the following experiment, it is fine-tuned using our Terahertz dataset. Another compared method, CGANs [15] models were trained using our Terahertz dataset with different Generators and loss functions, including the proposed Mask-CGANs, without any pre-trained network.

4.4. Experimental evaluation

Fig. 5 shows some samples segmented by our model. For visualization, we display the original Terahertz on the left, put ground

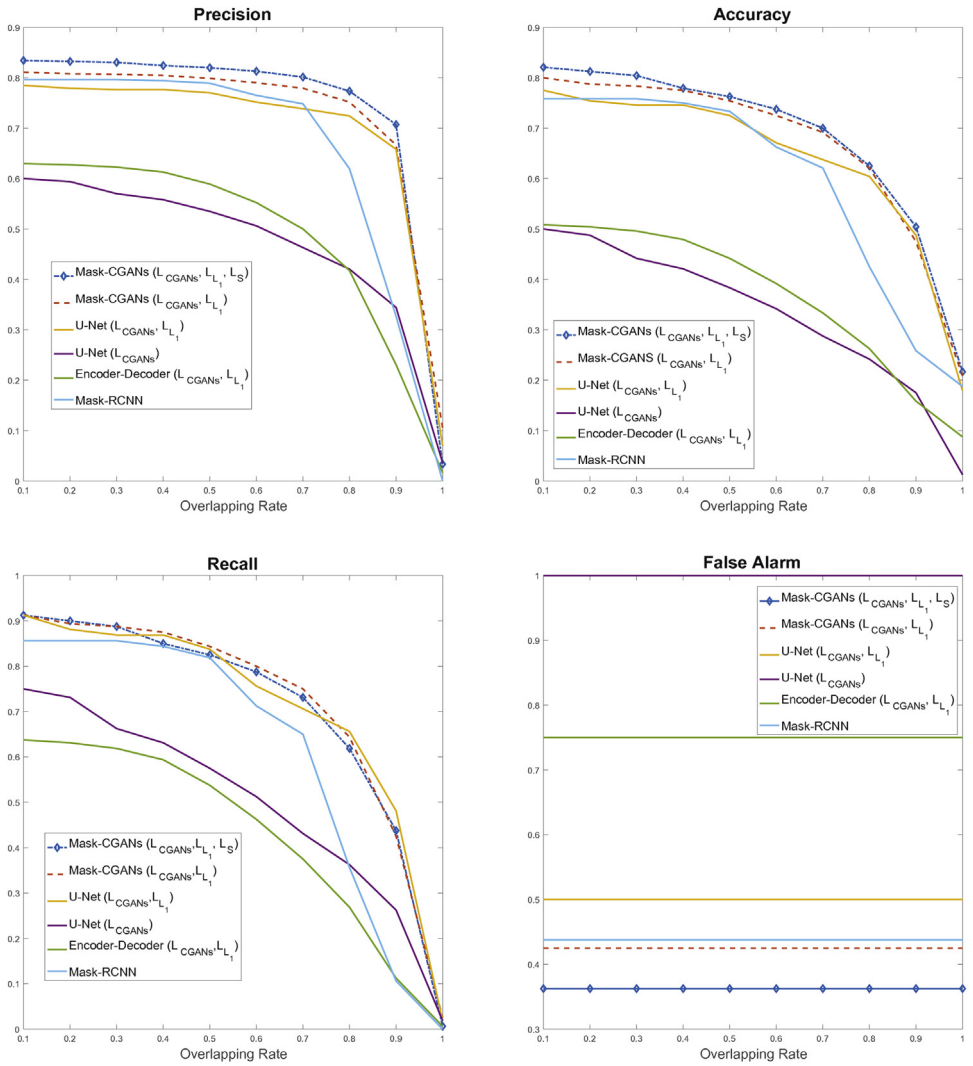

Fig. 6. The Precision and Accuracy, Recall and False Alarm rates are changed with overlapping area threshold changing.

truth (green) onto the sample in the middle and segmentation result (red) with the sample and the ground truth on the right. Our model can segment concealed weapons in noisy Terahertz samples where weapons are hidden at different locations by different persons with different postures. When the sample contains hidden weapons, our model can output a mask to cover its area, and the model outputs a black background for the samples without any concealed object.

The quantitative results in Table 1 have shown the efficiency of our Mask-CGANs model with appropriate loss function and model structure in this task. Without the L_{L_1} constrain ("U-Net"), which is a strong guidance to reach one-to-one match between the input and the output, the generation process is able to achieve suitable results for the segmentation task and its Precision and Accuracy rates are cut down much with a high False Alarm rate. For the model without concatenating low-level features ("Encoder-Decoder (L_{CGANs}, L_{L_1})") which play important role in finding tiny objects, its Recall rate is affected greatly (0.6375 vs. 0.9125 ("U-Net (L_{CGANs}, L_{L_1})"). However, our model structure is different from "U-Net" and gets good tradeoff between Recall and False Alarm rates in contrast to the other two candidates (False Alarm rate: 0.425 vs. 0.5 and 0.75). For sparse priors, the False Alarm rate decreased sharply (0.3625 vs. 0.4250), and Precision and Accuracy increases correspondingly.

Fig. 6 displays some comparison results, about Recall, Precision, Accuracy and False Alarm rates of different models, as the overlapping threshold changes. Because of our strict judgement method, the False Alarm rate will not be changed whatever the threshold is. Our final model Mask-CGANs with selected-connection U-Net and Loss Function (L_{CGANs}, L_{L_1}, L_S) outperforms others both in Precision and Accuracy rates at each threshold, even though Recall rates are affected slightly at some threshold points.

As for Mask-RCNN, our model surpasses this state-of-the-art instance segmentation method in such small and noisy specific dataset both in the indexes shown in Table 1 and segmentation testing speed. As shown in Fig. 7, our model's testing speed is **69.7** *FPS*, while Mask-RCNN's is **1.6** *FPS*. Note that the significant advantage in the processing speed makes it possible to be implemented in a real-time security inspection system.

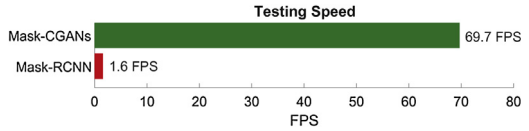

Fig. 7. The testing speeds of Mask-RCNN and our Mask-CGANs in this task.

Our model is trained for about 7.6 h in 400 epochs with 8 samples as a mini batch (all samples are resized to 256 × 256). All the experiments are tested on a single GTX1080G1 GPU within a Windows 10 operating system and TensorFlow.

4.5. Test on INBreast dataset

In this part, we show experiments on another dataset INBreast dataset [31], which is a mammographic mass analysis dataset, providing accurate contours of lesion region and the mammograms are of high quality. For mass segmentation, the dataset contains 116 mass regions. We use the first 58 masses for fine-tuning the model and the rest for test. For consistent medical comparison, the Dice index metric is used for the segmentation results and is defined as

$$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{9}$$

As shown in Table 2, the Dice index of the proposed method is 91.29%, surpasses the-state-of-the-art medical issue segmentation methods: Mask-RCNN [19] 91.07%, Multi-FCN-CRF with Adversarial Training [32] 90.97%, FCN with Adversarial Training 89.71%, FCN [33] 89.48%, CGANs with U-net [15] 90.40%, and U-net [28] 89.79%.

4.6. Interface implementation

The graphic user interface is built to show the experimental results more intuitively. The GUI is based on PythonQt as shown in Fig. 8. It displays the original Terahertz images (two images for one subject in fore and back view), the segmented mask of the concealed objects, and the bounding box based on the mask (the red bounding box added on the original Terahertz image).

4.7. Discussion

4.7.1. Dataset size and overfitting

A concern is that the dataset size is rather small compared with some deep learning standard datasets. Actually, the class number in our dataset is also small (1440 samples with 2 classes). Some deep learning datasets usually has samples more than 1000 in one class, but the number of the training samples is not the essence, the representativeness of data is one of the most important concerns for a machine learning task. Also, we applied data augmentation approaches, shift, rotation, deformations and gray value variations, in the training phase for each comparison method, which increased the number of the training data.

To test the generalisation of the proposed method, even our dataset is small, we still decide to use 1/3 independent data for testing. This rate is much larger compared with the traditional 10 folders or 5 folders testing.

GANs itself is a better choice when the training data is of a small number, that is also a consideration when we design this framework.

In summary, our model is trained with random data augmentation under an adversarial learning framework, and tested using a relative large independent testing dataset, to grantee the generalisation of this method.

4.7.2. False Alarm rate

Although the False Alarm rate of the proposed method is better than that of any other methods in our comparison, it is not so low due to our strict evaluation methodology [29] that one single positive pixel will result a false alarm sample. Our model is a generative method and some sparse local noise may not be fully removed. This kind of false alarm samples can be removed through a simple post-process like median filtering.

Table 2
Test on INBreast dataset.

Methods	Dice index
the proposed	**91.29**%
Mask-RCNN [19]	91.07%
Multi-FCN-CRF with Adversarial Training [32]	90.97%
FCN with Adversarial Training	89.71%
FCN [33]	89.48%
CGANs with U-net [15]	90.40%
U-net [28]	89.79%

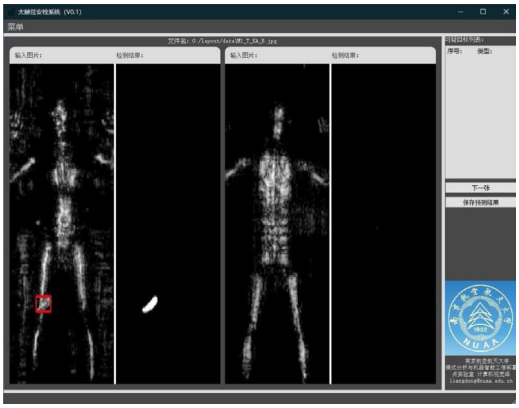

Fig. 8. The graphic user interface based on PythonQt, displays the original Terahertz images (fore and back view), the segmented mask, and the bounding box based on the mask (red bounding box added on the original image). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions and future work

We have presented a method to segment concealed weapon in Terahertz samples with poor imaging quality. Mask-CGANs with an optimal model structure and a proper loss function, has abilities in segmenting concealed objects in such low quality and noisy Terahertz samples within a small training dataset. The experimental results also prove that our system achieved much shorter processing time than Mask-RCNN. The fast processing speed means it is appropriate to be implemented in a real-time Terahertz security inspection system.

Mask-CGANs model still has a great improvement space. Improvements can be achievable in a more precise model with object class labels and then extend it to large-scale instance segmentation. Also, it provides a promising solution for other complex and low quality electromagnetic imaging environments, such as medical sample segmentation.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant 2017YFB0802300, National Natural Science Foundation of China61601223, Natural Science Foundation of Jiangsu ProvinceBK20150756, Postdoctoral Science Foundation of China (Top level)2015M580427. H. Zhou was supported by UK EPSRC under Grant EP/N011074/1, Royal Society-Newton Advanced Fellowship under Grant NA160342, and European Union's Horizon 2020 research and innovation program under the Marie-Sklodowska-Curie grant agreement No 720325.

References

[1] M. Kowalski, M. Kastek, M. Walczakowski, M. Szustakowski, N. Palka, Passive imaging of concealed objects in terahertz and long-wavelength infrared, Appl. Optics 54 (17) (2015) 3826–3833.
[2] K.B. Cooper, R.J. Dengler, N. Llombart, T. Bryllert, G. Chattopadhyay, I. Mehdi, P.H. Siegel, An approach for sub-second imaging of concealed objects using terahertz (THz) radar, J. Infrared Millim. Terahertz Waves 30 (12) (2009) 1297–1307.
[3] X. Yan, L. Liang, J. Yang, W. Liu, X. Ding, D. Xu, Y. Zhang, T.J. Cui, J. Yao, Broadband, wide-angle, low-scattering terahertz wave by a flexible 2-bit coding metasurface, Optics Express 23 (22) (2015) 29128–29137.
[4] J. Dai, K. He, J. Sun, Instance-aware semantic segmentation via multi-task network cascades, Comput. Vision Pattern Recognit. (2016) 3150–3158.
[5] P.H.O. Pinheiro, T. Lin, R. Collobert, P. Dollar, Learning to refine object segments, Eur. Conf. Comput. Vision 9905 (2016) 75–91.
[6] J. Dai, K. He, Y. Li, S. Ren, J. Sun, Instance-sensitive fully convolutional networks, Eur. Conf. Comput. Vision (2016) 534–549.
[7] J.R.R. Uijlings, K.E.A.V. De Sande, T. Gevers, A.W.M. Smeulders, Selective search for object recognition, Int. J. Comput. Vision 104 (2) (2013) 154–171.
[8] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping, Computer Vision and Pattern Recognition, (2014), pp. 328–335.
[9] P.H.O. Pinheiro, R. Collobert, P. Dollar, Learning to segment object candidates, Neural Inf. Process. Syst. (2015) 1990–1998.
[10] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, C. Rother, Instancecut: from edges to instances with multicut, Comput. Vision Pattern Recognit. (2017) 7322–7331.
[11] M. Bai, R. Urtasun, Deep watershed transform for instance segmentation, Comput. Vision Pattern Recognit. (2017) 2858–2866.
[12] A. Arnab, P.H.S. Torr, Pixelwise instance segmentation with a dynamically instantiated network, Comput. Vision Pattern Recognit. (2017) 879–888.
[13] S. Liu, J. Jia, S. Fidler, R. Urtasun, Sgn, Sequential grouping networks for instance segmentation, International Conference on Computer Vision (2017) 3516–3524.
[14] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Neural Information Processing Systems, (2014), pp. 2672–2680.
[15] P. Isola, J. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, Comput. Vision Pattern Recognit. (2017) 5967–5976.
[16] X. Shen, C. Dietlein, E.N. Grossman, Z. Popovic, F.G. Meyer, Detection and segmentation of concealed objects in terahertz images, IEEE Trans. Image Process. 17 (12) (2008) 2465–2475.
[17] D. Lee, S. Yeom, J. Son, S. Kim, Automatic image segmentation for concealed object detection using the expectation-maximization algorithm, Optics Express 18 (10) (2010) 10659–10667.
[18] D.S. Lee, J.Y. Son, M.K. Jung, S.W. Jung, S.J. Lee, S. Yeom, Y.S. Jang, Real-time outdoor concealed-object detection with passive millimeter wave imaging, Optics Express 19 (3) (2011) 2530–2536.

[19] K. He, G. Gkioxari, P. Dollar, R.B. Girshick, Mask r-cnn, Int. Conf. Comput. Vision (2017) 2980–2988.
[20] R.B. Girshick, Fast r-cnn, Int. Conf. Comput. Vision (2015) 1440–1448.
[21] S. Ren, K. He, R.B. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149.
[22] M. Mirza, S. Osindero, Conditional generative adversarial nets, Comput. Sci. (2014) 2672–2680.
[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, International Conference on Machine Learning (2016) 1060–1069.
[24] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C.S. Regazzoni, N. Sebe, Abnormal event detection in videos using generative adversarial nets, International Conference on Image Processing (2017) 1–5.
[25] X. Wang, A. Gupta, Generative image modeling using style and structure adversarial networks, Eur. Conf. Comput. Vision (2016) 318–335.
[26] C. Li, M. Wand, Precomputed real-time texture synthesis with markovian generative adversarial networks, Eur. Conf. Comput. Vision (2016) 702–716.
[27] G.E. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[28] O. Ronneberger, P. Fischer, T. Brox, U-net. Convolutional networks for biomedical image segmentation, Med. Image Comput. Comput. Assist. Interv. (2015) 234–241.
[29] W. Li, V. Mahadevan, N. Vasconcelos, Anomaly detection and localization in crowded scenes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (1) (2014) 18–32.
[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Comput. Vision Pattern Recognit. (2016) 770–778.
[31] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, Inbreast: toward a full-field digital mammographic database, Acad. Radiol. 19 (2) (2012) 236–248.
[32] W. Zhu, X. Xiang, T.D. Tran, X. Xie, Adversarial Deep Structural Networks for Mammographic Mass Segmentation, (2017) arXiv.org.
[33] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2014) 640–651.