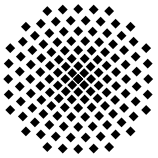


University of Stuttgart
Institute for Signal Processing and System Theory
Professor Dr.-Ing. B. Yang

Masterarbeit D1513

Range-Doppler map upsampling for single channel chirp sequence radar using Deep Learning

Upsampling der Range-Doppler-Karte für einkanaliges Chirp-Sequence-Radar unter Verwendung von Deep Learning

Author: Zheming Yin

Date of work begin: 01.10.2024
Date of submission: 31.03.2025

Supervisor: Sven Hinderer

Keywords: Keyword1, Keyword2 TBD

Contents

Abbreviations

CFAR	Constant false alarm rate
cGAN	Conditional Generative Adversarial Networks
CMGAN	Conformer-Based Metric-GAN
CNN	Convolutional Neural Network
CPI	Coherent Processing Interval
DP	Dual-Path
FFT	Fast Fourier Transform
FM	Frequency Modulated
FMCW	Frequency Modulated Continuous Wave
GAN	Generative Adversarial Networks
HQ	High-Quality
IF	Intermediate Frequency
iFFT	Inverse Fast Fourier Transform
LSD	Logarithmic Spectral Distance
LQ	Low-Quality
MIMO	Multiple input, multiple output
MSE	Mean Square Error
PLSD	Phase-aware Logarithmic Spectral Distance
PMSE	Pixel-wise Mean Squared Error
PSNR	Peak Signal-Noise Ratio
RA	Range-Azimuth
riFFT	Real-valued Inverse Fast Fourier Transform
RNN	Recurrent Neural Networks
SAR	Synthetic Aperture Radar
SDR	Signal-to-Distortion Ratio
STFT	Short-Time Fourier Transform
SDR	Scale-Dependent Signal-to-Distortion Ratio
Swin	Shifted Windows
SwinIR	Image Restoration using Swin Transformer
TF	Time-Frequency
TFRecord	TensorFlow Records

Abstract

Kurzfassung

1. Introduction

1.1. Chirp sequence radar

Chirp Sequence Radar is a radar system based on Frequency Modulated Continuous Wave (FMCW) technology, widely used in the field of automotive radar, target detection, and environmental sensing and so on. Chirp sequence radar transmits multiple consecutive chirp signals, receives the echoes, and calculates target information such as range and velocity, which can be used to generate a range-Doppler map.

FMCW radar

In FMCW radar, the chirp frequency increases linearly with time, of which the signal is called the linear Frequency Modulated (FM) signal. Figure 1.1 illustrates the change of a linear FM chirp in amplitude and frequency over time. The signal frequency will increase from the initial value f_c by the bandwidth B during the duration of one chirp T_c. After an time interval, the other chirp will be transmitted at time T_p. The rate of the frequency change is shown as the slope, namely $S = B/T_c$.
With this property, we can get the range and velocity of the other targets related to the FMCW radar. There will be a delay between the signal being transmitted and being received by the radar, which is proportional to the range. Since the signal is a round trip motion, the range between the radar and the target can be defined as

$$r = c \cdot \frac{\tau}{2}, \tag{1.1}$$

where c represents the speed of light, namely 3×10^8 m/s, and τ represents the propagation delay. The Doppler effect refers to the phenomenon that the frequency of the received signal and the transmitted signal will be different when there is a relative velocity between the signal source and the target. The relationship between the frequency of the transmitted signal f and the received signal f' is

$$f' = f\left(\frac{c \pm v_o}{c \mp v_s}\right), \tag{1.2}$$

where v_0 and v_s represent the relative velocity of the radar and target to the medium, respectively. When they move towards each other, the frequency of the received signal increases, and vice versa. In the collected dataset, in most cases the radar as the signal source is moving, while the surrounding objects are stationary.
Meanwhile, as the received signal is captured by the radar, it will be mixed with the transmitted signal in the system to obtain a new signal called Intermediate Frequency (IF) signal. The frequency of this IF signal is proportional to the delay and propagation distance, written as

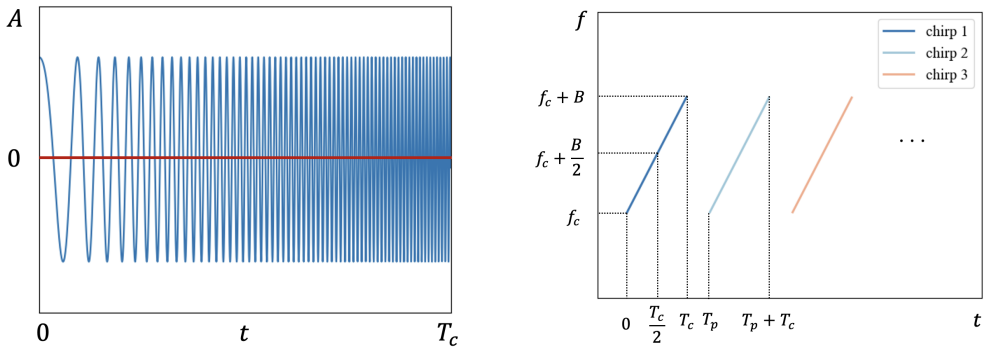


Figure 1.1.: The amplitude and frequency variation of FMCW radar in time respectively

$$f_{\text{IF}} = S \cdot \frac{2r}{c}. \tag{1.3}$$

Chirp sequence radar

The chirp sequence radar is a FMCW radar system with the steep frequency ramps [**fink_comparison_2015**]. The dataset in this paper are collected from a single channel chirp sequence radar, which means only a single transmitter and a single receiver were used. In section 2.1, the hardware of the chirp sequence radar will be explained in detail, where there are three transmitters and single receiver. The advantages of using a single transmitter are following, on the one hand, it mitigates the interference problem. When multiple transmitters work at the same time, the radar system will experience beat frequency aliasing, which will cause errors in the results. And if there are phases between different transmitters, the accuracy of Doppler measurement will also be reduced. On the other hand, the signal needs to do the Fast Fourier Transform (FFT) to convert the signal from time domain into the frequency domain, and the computational complexity will increase accordingly. Meanwhile, the Multiple input, multiple output (MIMO) radar will cause a higher cost.
In addition to the difference in the rate of frequency change, another advantage of chirp sequence radar is that it can achieve a larger bandwidth. In section 2.3, the formula of range resolution will be derived, in which the bandwidth and range resolution are inversely proportional, therefore, a larger range resolution can be achieved.

Range-Doppler map

According to the above, the chirp sequence radar based on the FMCW radar can also obtain information about range and velocity. Meanwhile, there is a corresponding relationship between the range and velocity, which can be drawn through the range-Doppler map. The derivation and visualization process will be shown in detail in section 2.3, but the range-Doppler map looks like the Figure 1.2.
In the range-Doppler map, the horizontal and vertical axes are range and velocity respectively. In our dataset, the range is calculated from 0 m to roughly 10 m, and the velocity changes roughly from -5 m/s to 5 m/s. At each point where the range and velocity intersect, there is a value representing the amplitude of the signal power information. When the amplitude is converted to dB unit, the value range is between -80 dB and 60 dB, and the corresponding color in the bar gradually changes from dark to light.

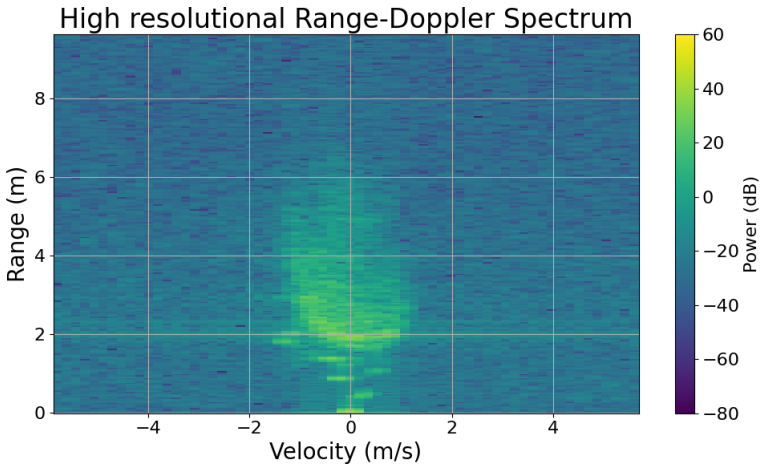
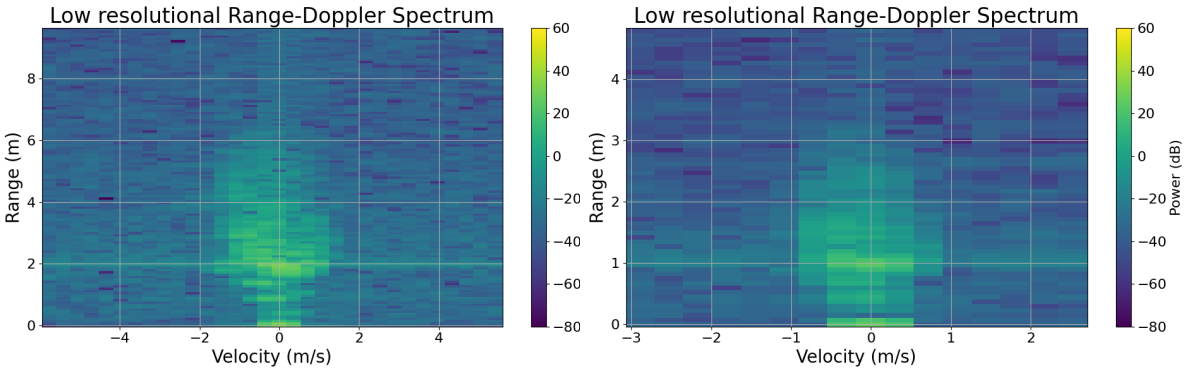


Figure 1.2.: An example of the range-Doppler map

1.2. Motivation

As mentioned above, the chirp sequence radar can achieve a larger bandwidth and higher range resolution, but in practical applications, many factors will limit the resolution, such as carrier frequency and Coherent Processing Interval (CPI). The carrier frequency is a fixed value in a radar system. When its frequency is not high enough, the wavelength is longer and the bandwidth becomes smaller, thereby reducing the range resolution. The carrier frequency used by the radar system is affected by many factors as well, such as regulation, and low carrier frequency radar systems can reduce costs. CPI refers to the time interval in which the radar continuously observes the target and performs signal processing. The signal during this time period is coherent, that is, the phase of the signal remains stable and does not change significantly. In this process, the longer the CPI, the more accurate the velocity can be estimated, since the radar can collect more data to reduce the background noise and improve the signal quality. However, in many cases, due to the movement of the detection targets or the ego-motion of the radar itself, the CPI is limited, which in turn affects the velocity resolution.
When Figure 1.2 is regarded as a high-resolutional range-Doppler map, then in reality, the resolution of range and velocity may be only half or even a quarter of that, resulting in a lot of information loss, as shown in Figures 1.3, where Figure 1.3a shows the effect of the range-Doppler map with the halved resolution, whereas Figure 1.3b shows a quarter resolutional image.
However, high resolutional range-Doppler maps are crucial and have a wide range of applications in areas such as the indoor localization. In order to accurately process and utilize the range-Doppler map, such as Constant false alarm rate (CFAR) detection, a high resolutional range-Doppler map is expected, and wish as many details as possible can be kept. Traditional methods such as the interpolation, which uses linear interpolation based on surrounding values to achieve upsampling. With the development of deep learning, it has shown a strong capability to upsample image to the high resolution. The commonly used structures and models include Convolutional Neural Network (CNN), Encoder-Decoder, Transformer, Conditional Generative Adversarial Networks (cGAN), Diffusion and so on. Meanwhile, there are also multiple super resolution solution, such as the super resolution system in video



(a) In halved resolution (b) In quarter resolution

Figure 1.3.: Examples of the low resolutional range-Doppler map in different downsampled factor

games provided by NVIDIA.
However, the current upsampling approaches are mainly based on images or videos generally, and rarely specifically based on the range-Doppler map. Compared with the daily images, the range-Doppler map has several significant characteristics. One is that the values in the range-Doppler map have relatively large fluctuations. According to Figure 1.2, the amplitude can be as high as 40dB in places with a closer range, while the noise will be lower than -60dB in places farther away from the radar. Furthermore, the values in the range-Doppler image are not as coherent and smooth as those in the images we usually see. In addition to the higher values in the parts with closer ranges, there are usually large differences between even two adjacent points in the surrounding area, which asks for the special requirements on data processing and structure of the model. We would like to process the data so that its fluctuation range is relatively small. Meanwhile, since each point in the range-Doppler map contains a variety of information such as velocity, range, power, angle and so on, there is a certain correlation between these information. We hope that the model can learn the connection and relationship between different areas and use deep learning to achieve better upsampling results than the traditional methods such as interpolation.
In addition, another challenge is that our goal is not just based on the closer part, i.e., the area with higher amplitude, but the entire range-Doppler map, including the lower amplitude signals or background noise. Therefore, in data processing, the purpose of making the range of amplitude smaller not only reduces the fluctuation range of the data, but also allows the lower amplitude signals to receive the same attention.
This thesis will mainly focus on three parts: dataset, models, and loss functions. We will use the radar in the ISS to collect the data to create range-Doppler maps and process the dataset. The pipeline will be built to load dataset, build and train models, and evaluate them. We will also design the loss function as well that meets our needs to achieve better upsampling on the range-Doppler map.

1.3. Structure of the thesis

The thesis will consist of seven chapters. Chapter 1 will focus on the introduction of the content in the title and the explanation of some basic concepts, which are completed in section

1.1. Section 1.2 introduces the reasons why this thesis will upsample based on the range-Doppler map, the challenges, and the ultimate goal of this thesis. Section 1.3 will show the main contents of the entire paper and each section. Section 1.4 will show the current upsampling approaches in the state of the art, including but not limited to upsampling models for range-Doppler maps.
Chapter 2 will focus on data part. Section 2.1 will introduce the information of data collection, including the hardware of the radar used, the parameters of radar initialization, the settings during the collection process, such as the motion and environmental conditions, and the size of the collected dataset. In addition, the resampling operation of the data will be demonstrated. Based on the collected data as the high-resolution image, i.e., the ground truth, it will be downsampled to the training dataset. Section 2.2 will show the process of loading the dataset, using TFRecord files to store data to speed up the loading and training process, and the setting of parameters while creating TFRecord files, such as the frame of length. Meanwhile, since the plotting of the range-Doppler map is on the frequency domain, the conversion between the time domain and the frequency domain for the range-Doppler map will also be demonstrated. Section 2.3 derives the maximum range, the maximum velocity, and the resolution of the range and velocity to plot the range-Doppler map. Section 2.4 will show how to process the model's inputs and outputs. Since the model can only process float value rather than complex value, it could be converted into the concatenation of the real and complex part or of the amplitude and phase part, and they can be also normalized to reduce the fluctuation.
Chapter 3 is mainly about the deep learning models used, including a relatively simple CNN model, a simple U-Net structure and a U-Net model that concatenates the encoder part and the decoder part, and two different architectures using transformers, called Dual-Path (DP)-Time-Frequency (TF) architecture and Image Restoration using Swin Transformer (SwinIR) structure, respectively in section 3.4 and 3.5. The transformer has also two types, DP block and Shifted Windows (Swin) block. Section 3.6 will compare the differences between the transformer architectures and blocks. Section 3.7 will use the above models to design a generator and introduce a discriminator to implement the adversarial model. Additionally, the model must undergo an upsampling layer before the output. There are currently two main methods, which will be introduced in section 3.8. In the process of converting data from the time domain to the frequency domain, the dimension will turn to an odd value, such as from 512 to 257, which will cause problems when the model divides the blocks. Therefore, section 3.9 will focus on this problem and propose two methods.
Chapter 4 focuses on various loss functions, and based on the effects of different loss functions, section 4.6 will combine them to take advantage of different pros and cons, then complement each other. Chapter 5 will show some methods used in the pipeline to optimize the entire training process, such as using multiple GPUs for distributed training to speed up the training process in section 5.1, and the hyper- parameters which can be tuned as well as the WandB sweep function shown in section 5.3.
Finally, Chapter 6 shows the impact of different processing and models on the upsampling results and analyzes them. Chapter 7 summarizes the results of the entire thesis and gives the outlook of the further research on the remaining deficiencies.

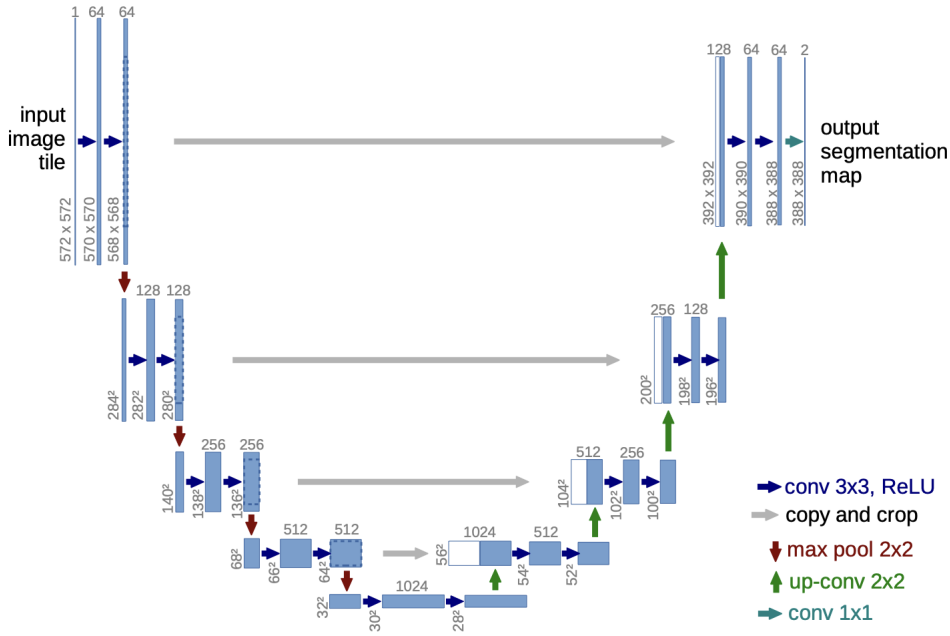


Figure 1.4.: U-Net architecture [**ronneberger_u-net_2015**]

1.4. Overview of the state of the art

The super-resolution methods of recent advances mainly focuse on the supervised learning algorithms, since it is hard to obtain the same images with different resolutions in the same scene [**wang_deep_2021**]. Currently, the models and structures used in super resolution mainly include CNN, residual-based methods, encoder-decoder structure, GAN model, etc [**li_review_2021**]. In addition, Transformer has strong ability in terms of translation task and wording understanding. The model uses Attention mechanism to understand the association and relationship between different words and phrases, so that it can understand longer sentences than Recurrent Neural Networks (RNN) [**vaswani_attention_2023**]. Taking advantage of this feature, the Swin Transfomer designs a shifted window based on Transformer divides the image blocks and shifts between them, then different blocks can also learn the relationship between each other [**liu_swin_2021**].

Encoder-decoder structure

The encoder-decoder includes two parts. First, it will use encoder to extract features, learn the extracted features, and then use the decoder part to restore the features, so that the model can get the required output. A common structure is called U-Net. Olaf et al. apply U-Net to segment medical images. According to Figure 1.4, the structure of their model presents in U-shaped. The left side of "U" is the encoder part, that is, the features of the image are extracted and learned, and the right side is the part of the decoder which uses the learned features to restore the image to its original size, while segmenting the objects in the image [**ronneberger_u-net_2015**].
In the field of image super-resolution, Akarsh et al. used U-Net to upsample the point clouds of mmWave radar. A structure similar to residual is used in the model, concating the decoder part with the corresponding position of the encoder to mitigate the information loss. The pixel-wise loss and dice loss are combined in the loss function to improve both pixel-wise

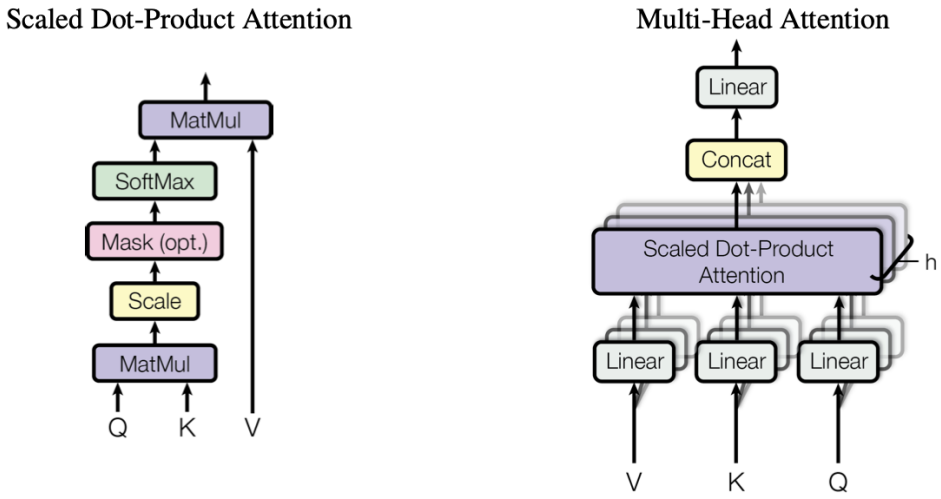


Figure 1.5.: Attention mechanism [**vaswani_attention_2023**]

accuracy and the clarity of the boundary. In the evaluation part, the paper used different environments to test the robustness of the model [**prabhakara_high_2023**]. Li et al. also applied U-Net to the Range-Azimuth (RA) map collected by FMCW radar with MIMO. They have much data processing, inclusive converting the data with the real and imaginary parts, using FFT to convert the data to a more informative map, and choosing the shuffle pixel method for upsampling.

Transformer & Swin Transformer

In Transformer, the encoder-decoder structure exists as well. There are six identical layers in the encoder part, each layer includes a multi-head attention and a forward propagation sub-layer. According to the attention layer shown in Figure 1.5, it will calculate the similarity between the query, keys and values, thereby obtaining the relationship between different words. In the image task, we will use the self-attention mechanism, that is, query, keys and values all come from the same image, thereby obtaining the relationship between the image windows [**vaswani_attention_2023**]. It is on this basis that the Swin transformer adds shifted windows to achieve that there can be overlapping between windows to improve the uncertainty brought about by fixed split [**liu_swin_2021**]. Liang et al. proposed the SwinIR architecture based on the Swin transformer, adding shallow feature extraction and High-Quality (HQ) image restoration. The Low-Quality (LQ) image will be downsampled once and then passed through multiple Swin transformers and residual blocks to improve the image resolution [**liang_swinir_2021**].
Sven proposed a noval approach to separate two blind signals, which also uses Transformer model, but split the signals along the two coordinates of time and frequency respectively. In addition, due to two signals, it uses a dual-path structure to achieve better distinction between different signal sources [**hinderer_blind_2022**]. Inspired by this separation method, since there are also two obvious dimensions in the range-Doppler map, namely the range and the Doppler effect, we can also try to apply Transformer along these two axes respectively, and the overlapping will exist in this split approach as well.

GAN & cGAN models

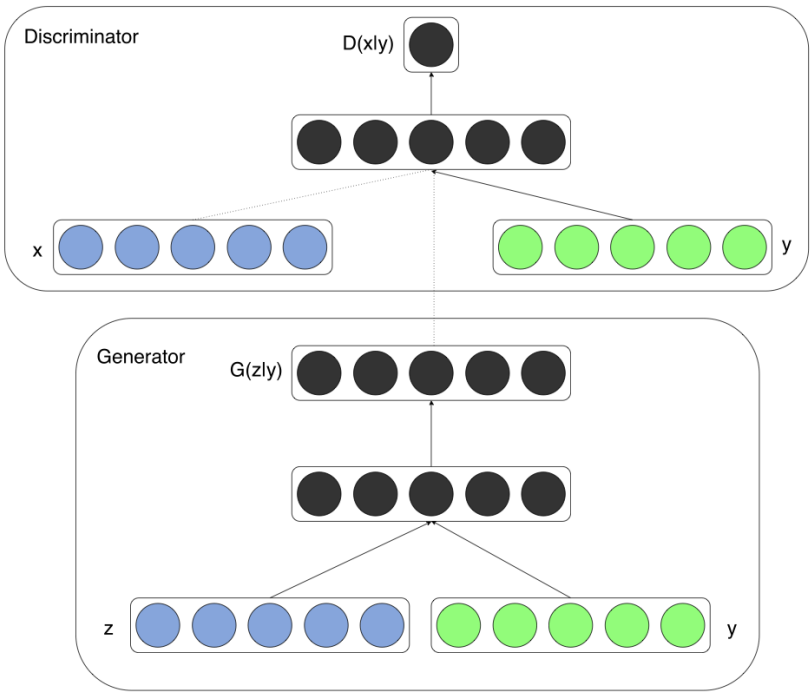


Figure 1.6.: cGAN architecture, where x is the ground truth, z represents the noise and y is the condition of the model [**mirza_conditional_2014**]

The Generative Adversarial Networks (GAN) model also contains two parts, called generator and discriminator. The generator part will extract and learn the data distribution of the input. After generating an output, the discriminator is used to estimate the probability that the ground truth and the generated output are respectively the real one. The goal is to make the output of generator as close to the ground truth as possible. On the other hand, it is hoped that the discriminator can enhance its discrimination ability and jointly improve the learning ability of the model through adversarial learning between the two parts [**goodfellow_generative_2014**].
Based on the Generative Adversarial Networks (GAN) model, Mehdi et al. proposed a conditional version of GAN model, called Conditional Generative Adversarial Networks (cGAN). cGAN will feed data in the generator and discriminator parts, which is called the condition, as the "y" demonstrated in the Figure 1. For image processing tasks, this condition is the image itself. In our case, the generator input will include the low resolutional range-Doppler map, and the discriminator will provide the corresponding high resolutional range-Doppler map [**mirza_conditional_2014**].
In the field of image super-resolution remedy, Karim et al. used the GAN model on the data of FMCW radar, which was based on the micro-Doppler signature of walking human targets. The GAN model consists of two parts: the discriminator and the CasNet generator, where the CasNet generator is a stack of multiple U-Nets. The loss function combines the pixel-wise L1 loss and the perceptual loss obtained by the discriminator [**armanious_adversarial_2019**]. Compared with the task in this thesis, the content of the range-Doppler map is more complicated and the fluctuation of the data is more severe.
With the usage of cGAN, Kong applied it to the super-resolution of Synthetic Aperture Radar (SAR) imagery. The structure of U-Net was used in the generator, downsampling was per-

formed through CNN and upsampling was performed through Transposed CNN, and an attention mechanism was performed after each dimension change. Furthermore, multiple losses were combined in the loss function, including GAN loss, VGG perceptual loss and feature matching loss [**kong_dmsc-gan_2023**]. Inspiration from this, multiple loss functions can also be combined in our training process to achieve better results. Sherif et al. used the Conformer-Based Metric-GAN (CMGAN) model, combining the Transformer with the GAN structure to learn long-distance dependencies and the CNN to exploit local features, which can tackle a variety of tasks, including super-resolution processing of speech source. The same Short-Time Fourier Transform (STFT) processing method was used as the DP-TF architecture above, and the real part, imaginary part and amplitude are concatenated as the input of the model [**abdulatif_cmgan_2024**].
Also in the field of Rang-Doppler map super-resolution, Jeong et al. used the cGAN architecture, in which the generator was based on the U-Net model. In the evaluation, more metrics are used, such as Pixel-wise Mean Squared Error (PMSE), Peak Signal-Noise Ratio (PSNR), etc [**jeong_resource-efficient_2023**]. However, since the data was collected mainly in open areas such as playground, the content in the Rang-Doppler map is relatively not that much and single. Ledig et al., in their SRGAN model, a GAN of the image super-resolution, for the first time used a larger resampling rate, namely factor of four, which put higher requirements on the ability of the model. Meanwhile, they used perceptual loss to combine adversarial loss and content loss to prevent the loss of high-frequency information and ensure that the boundaries of objects are clearer [**ledig_photo-realistic_2017**].

2. Dataset

2.1. Dataset recording

The hardware of the radar product, the parameters, settings of recording, the environment and environment split, and the size of the dataset as well as resampling.

2.2. Dataset loading

the settings & Creation of the TensorFlow Records (TFRecord) files creation such as frame of length, shift window for augmentation, cube de-biasing, iFFT, FFT shift, riFFT.

2.3. Range-Doppler map visualization

plotting the rD map in logarithm, resolution calculation, maximum range and velocity, limit the value range of the bar.

2.4. Pre- and post-processing of the model inputs and outputs

2.4.1. Real/Imaginary separation

2.4.2. Amplitude/Phase separation

incl. logarithm operation & 10.0 power and cut-off to avoid the nan problem

2.4.3. Input normalization

the normalisation of the input to be (-1,1) or (0,1), and the global min/max/mean

3. Models

3.1. Interpolation

3.2. CNN model

3.3. U-Net architecture

3.3.1. U-Net simple

3.3.2. U-Net concat

3.4. DP-TF Transformer architecture

3.5. SwinIR Transformer architecture

3.5.1. DP Transformer block

3.5.2. Swin Transformer block

3.6. Comparison between the architectures

3.7. cGAN architecture

3.7.1. Generator

3.7.2. Discriminator

3.8. Upsampling layer

Transposed convolutional layer or Shuffle pixel

3.9. Dimension processing

padding or convolutional layer to make (129,32) to be (130, 32) or (128,32)

4. Loss functions

4.1. MSE loss

show the MSE & weighted MSE

4.2. SDR loss

show the SDR & SD-SDR

4.3. LSD loss

show the LSD and PLSD

4.4. Perceptual loss

show the different VGG layer

4.5. GAN loss

show both generator loss and discriminator loss

4.6. Combination of the losses

5. Training optimization

5.1. Distributed training

5.2. Settings in the pipeline

5.2.1. Early stopping

5.2.2. Learning rate schedule

5.2.3. Checkpoint setting

5.2.4. Optimizer

5.2.5. Pre-training

5.2.6. Overfitting attempt

5.3. Hyper-parameter optimization

which parameters have been tuned, manual tuning and tune by wandb sweep

6. Results

The following evaluations can use the metrics, such as MSE, LSD, SDR, VGG perceptual loss in both frequency and time domain as well as plotting the super resolutional result.

Data processing

1) Real/imaginary & Amplitude/phase separation & only amplitude as the inputs of the model
2) Influence of the logarithm operation & log10 or log2
3) Normalization & the range, such as (-1,1) or (0,1) for amplitude and (-pi,pi) or (-1,1) for angle

Model

4) The upsampling layer: Transposed conv or shuffle pixel
5) Dimension processing: padding or conv layer
6) Comparison between different models (CNN, U-Net, DP, SwinIR)
7) Influence of the differences between DP and SwinIR architectures and Transformer blocks
8) Number of the parameters in models
9) Influence of the discriminator
10) The input of the discriminator, with or without low resolutional input
11) Activation function of the last conv layer in models (no act / cut-off / sigmoid if in the range of (0,1))

Loss function

12) Different loss functions in the training process & combination of the loss functions
13) Mask in the LSD loss function
14) Combination of the different VGG layers in the perceptual loss

Other settings

15) Influence of the pre-training in the cGAN
16) Resampling rate as factor 2 & factor 4 for the low resolutional image
17) Frame of length, 1 or 2 or 4
18) Training speed with distributed training
19) Hyper-parameters tuning (such as learning rate, lambda in loss function and phase loss and so on)

7. Summary & Outlook

7.1. Summary

7.2. Outlook

A. Appendix

You may do an appendix

List of Figures

List of Tables

Declaration

Herewith, I declare that I have developed and written the enclosed thesis entirely by myself and that I have not used sources or means except those declared.

This thesis has not been submitted to any other authority to achieve an academic grading and has not been published elsewhere.

Stuttgart, TBD Date of sign. Zheming Yin