Efficient Network Design for High Dimensional Data

by

XIN MIAO

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2020

To my family

ACKNOWLEDGEMENTS

I was born in a very small city in China. Before I was 20 years old, if someone told me that I would get a Ph.D. in the United States in the future, I must think it was a joke. I never thought I would come to study and work on the other side of the ocean. So I am very happy that I got this Ph.D. It is a very meaningful journey in my life, I have not only learned a lot, but also met many interesting friends.

First of all, I would like to express my deepest gratitude to my supervior, Dr. Vassilis Athitsos. He is a very patient and kind person. He gave me a lot of guidance in scientific research. During my PhD time, I had a very pleasant time.

Secondly, I would like to thank my committee members, Dr. Christopher Conly, Dr. Dajiang Zhu and Dr. Farhad Kamangar, for all the assistance on this dessertation.

Thirdly, thank you to my girlfriend Junjing Zhang, for all her love and support. I can't imagine how would be without her. Looking forward to the near future, we can form a happy family and have cute children, lovely dogs and cats. And my dream is to be able to retire in my forties, take her to travel around the world and walk through every interesting corner of the world.

Lastly but most importantly, my parents, Mr. Yanli Miao and Mrs. Zongqiao Chang. I am so grateful to have them in my life. In fact, my parents encouraged me to study in the United States. Although they will miss me very much, they hope I can achieve greater success.

Getting a doctorate is the biggest achievement in my life so far, I hope it can bring more help to my life, and I can enjoy life better because of it

April 15, 2020

ABSTRACT

Efficient Network Design for High Dimensional Data

XIN MIAO, Ph.D.

The University of Texas at Arlington, 2020

Supervising Professor: Vassilis Athitsos

Due to the powerful feature representation capabilities, deep learning has became a powerful tool in the field of computer vision. Especially in the aspect of high-dimensional images, deep learning can achieve fast inference compared with most traditional methods. This paper focuses on how to design an efficient neural network and apply it to two high-dimensional images application, video facial landmarks detections and compressive imaging system.

In this first part of this paper, we focus on landmarks detection for video facial images. Existing methods for facial landmarks detection mainly rely on cascaded regression. It is an indirect method and progressively estimates shape increments in an iterative way. Moreover, cascaded models extract handcrafted features, which fail to leverage the strength of convolutional neural networks. In addition, those local descriptors need to be calculated in each iteration based on updated shapes, which can be time consuming and makes it hard to integrate feature learning into one single architecture for end-to-end learning. This paper propose the a direct shape regression network (DSRN) which can achieve fast facial landmarks prediction. Specifically, by deploying doubly convolutional layer and by using the Fourier feature pooling layer proposed in this paper, DSRN efficiently constructs strong

representations to disentangle highly nonlinear relationships between images and shapes. It can run very fast with about 500 frames per second excluding face detection in the platform of NVIDIA GTX 1080Ti GPU, which is promising for the prospect of practical video facial landmarks detection.

In this second part of this paper, we proposes a deep learning framework for high dimensional images reconstruction in the snapshot compressive imaging system. Snapshot compressive imaging (SCI) refers to compressive imaging systems where multiple frames are mapped into a single measurement, video compressive imaging and hyperspectral compressive imaging are two representative aspects. In this manner, a two-dimensional (2D) monochromatic camera can sample the scenes at video rate and thus saves memory, bandwidth and cost significantly. While enjoying all these advantages, one important step in SCI is that algorithms are required to reconstruct the 3D data-cube from every snapshot measurement after the sensing process. Existing algorithms are either too slow or the performance is not high which preclude wide applications of SCI. In this paper, we develop a dual-stage deep learning model to reconstruct the desired 3D signal in SCI. It can be used for both video reconstruction and hyperspectral images reconstruction. The only difference is just the training process for the deep network. Results on both simulation and real datasets demonstrate the significant advantages of our network, which leads to a huge improvement in PSNR on simulation data compared to the current state-of-the-art. Furthermore, our network can finish the reconstruction task within sub-seconds instead of hours taken by the most recently proposed DeSCI algorithm, thus speeding up the reconstruction >1000 times.

Keyword: Deep Learning; High Dimensional Images; Snapshot Compressive Imaging; Facial Landmarks Detection

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

Introduction

In computer vision, we usually need to process high dimensional data. For example, there are a lot of methods are designed for video analysis application like action recognition [1, 2, 3, 4, 5, 6, 7, 8], semantic segmentation [9, 10], medical imaging [11, 12, 13, 14, 15] and vehicle identification [16]. Also some quantization methods were proposed to improve the speed of the network [17, 18]. Facial landmarks detection is an application which need the model can do fast inference during the testing. However, most existing methods can not satisfy this requirement. Another example for high dimensional data processing is snapshot compressive imaging (SCI). It refers to compressive imaging systems where multiple frames are mapped into a single measurement, video compressive imaging and hyperspectral compressive imaging are two representative aspects. However, due to the complicted process of the most of algorithms and limitation of the computational resource, it is not easy to apply a fast method for a real high dimensional data application. In the first part of this paper, we use deep learning method to solve one of the video analysis task. We propose the a direct shape regression deep network for end-to-end face alignment without relying on popular cascaded regression. It can run very fast with about 500 frames per second excluding face detection in the platform of NVIDIA GTX 1080Ti GPU, which is promising for the prospect of practical video facial landmarks detection. In the second part of this paper, we propose an efficient deep netwok which can be used in the compressive imaging and can do fast images reconstruction [19, 20].

1.1 Deep Learning for Facial Landmarks Detection

Facial landmarks detection has recently drawn great popularity in computer vision due to its prerequisite role in facial image analysis e.g. face recognition [21], face verification [22] and facial attribute analysis [23]. Facial landmarks detection is to estimate a set of predefined key points, which is known as landmarks providing semantic description of facial shapes. Facial landmarks detection has been studied extensively in recent years, while it remains an outstanding task. Its great challenges stem from the nonlinear relationship between input images and output shapes, since images are usually represented by low-level features while facial shapes contain high-level semantic meanings. Meanwhile, landmarks are spatially correlated, which can also be exploited for more robust and accurate alignment.

Cascaded regression has dominated in facial landmark detection and made great progress in the past decades. Nevertheless, the cascaded regression model suffers from intrinsic shortcomings. It is indeed an indirect method and progressively estimates shape increments in an iterative way, which highly depends on initialization. Therefore, the final solution of cascade models would likely to be trapped in local optima if the initialized shape is far from the true shape. Cascade models rely on local feature descriptors, and only the regions around landmarks are passed through the feature extractor. As a result, the smantic information of faces and correlations between landmarks are largely overlooked. Moreover, cascaded models extract handcrafted features, e.g., SIFT [24], HoGs [25], which fail to leverage the strength of convolutional neural networks. In addition, those local descriptors need to be calculated in each iteration based on updated shapes, which however can be time-consuming and would not be easily integrated in one single architecture for end-to-end learning. Thus, it is not easy to apply the cascaded regression for the video based facial landmarks detection which require the real time prediction.

In this paper, we propose direct shape regression networks (DRSN) to directly predict facial landmarks from images without relying on cascaded regression. DSRN tackles the aforementioned challenges by jointly modeling input-output relationships and landmark correlations in compact end-to-end learning architecture which composed of doubly convolutional, Fourier embedding and low-rank learning layers.

DRSN incorporates the doubly convolutional module to fulfill feature extraction [26], which is computationally more efficient due to less parameters while improving performance compared to regular convolution. It well fits face alignment tasks where training samples are rather limited compared to other vision tasks, e.g., image classification. In conjunction with the doubly convolutional module, DSRN introduces Fourier feature embedding into the last convolutional layer to build strong holistic representations. The Fourier embedding is derived from kernel approximation to leverage the strong ability of kernel methods for nonlinear feature extraction [27], which enables handling the nonlinear relationship between images and shapes. As a consequence, Fourier embedding accomplishes a nonlinear layer with a cosine activation function, which is readily learned in an end-to-end way by back-propagation.

Meanwhile, previous cascaded face landmarks detection has largely overlooked the landmark correlation and not yet been modeled explicitly. By properly exploiting the correlation, it can not only help to recover occluded landmarks but also improve the overall estimation performance. We propose encoding the correlation in a principled way. Specifically, we design a new linear layer with a bottleneck structure by low-rank factorization to replace the fully connected layer as the output layer. The low-rank learning is able to explicitly encode the intrinsic correlations by forcing correlated outputs to share subsets of features due to the low-rank factorization. More importantly, the low-rank layer can efficiently learned due to its nature of linearity with suffering from bad local minimum.

In this paper, our major contributions can be summarized in the following three aspects.

- We propose a direct shape regression model for end-to-end face landmarks detection without relying on cascaded regression. Our method accomplishes a novel compact convolutional learning architecture, which leverages the strengths of kernel methods for nonlinear feature extraction and convolutional neural networks for multivariate structured prediction.
- We propose a new feature extraction layer which is composed of a doubly convolutional layer and a Fourier feature embedding layer to efficiently build strong feature representations, which enable disentangling the highly nonlinear relationship between images and the associated shape of facial landmarks.
- Our model can run very fast with about 500 frames per second excluding face detection in the platform of NVIDIA GTX 1080Ti GPU, which is promising for the prospect of practical video facial landmarks detection.

The great effectiveness of the proposed DSRN has been verified exhaustively on four benchmark static datasets including CelebA, MAFL, ALFW, 300-W and one video based dataset 300-VW. Experimental results show that DSRN consistently achieves high estimation accuracy on all datasets and produces new state-of-the-art performance, largely surpassing previous methods by up to 30%. In contrast to cascaded models, once learned in the training stage, DSRN can efficiently predict landmarks on new input face image by simple matrix multiplications without further iterative optimization. More importantly, our DSRN offers a general compact convolutional learning architecture for face alignment, which can be readily used for real time video facial landmarks detection.

1.2 Deep Learning for Compressive Imaging Systems

Snapshot compressive imaging (SCI) refers to compressive imaging systems where multiple frames are mapped into a single measurement, video compressive imaging and hyperspectral compressive imaging are two representative aspects. The first SCI system, called coded aperture snapshot spectral imaging (CASSI), was developed in [28], which modulates signals at different wavelengths by a coded aperture (physical mask) and a disperser [29]. In this manner, a two-dimensional (2D) monochromatic camera can sample the hyperspectral scenes at video rate [30] and thus saves memory, bandwidth and cost significantly compared with that using a traditional spectrometer in addition to the high-speed sensing. While enjoying all these advantages, similar to other computational imaging systems, one important step in SCI is that algorithms are required to reconstruct the 3D datacube from every snapshot measurement after the sensing process. Existing algorithms are either too slow or the performance is not high.

Though deep learning based algorithms have started being used in computational imaging systems [31, 32], significant challenges and questions exist in SCI reconstruction using deep learning.

1) Limited training dataset is available. Firstly, the hyperspectral image itself is a high-dimensional cube. Secondly, though some datasets [33, 34] can be downloaded, the spectral wavelengths are usually different for different imaging systems. In order to overcome this challenge, in addition to the generally used data argumentation techniques, rotating and flipping, we further use the spectral interpolation to unify the datasets to the same set of wavelengths.
2) The measurement of SCI is a single frame, while more than 20 spectral channels (24 is used in our hyperspectral experiments) are to be generated (reconstructed). Therefore, a deep generative model has to be used. It has been observed that for compressive imaging reconstruction, a deeper generative model usually leads to better results [32]. However,

this is challenging due to the large number of parameters in the network and also the limited dataset mentioned above.

3) The third question this paper aims to address is that is it possible to adopt a small network to boost up the quality of SCI reconstruction results?

Bearing these challenges and questions in mind, this paper makes the following contributions [35, 36, 37, 38].

- A generative model based on U-net [39] is developed to reconstruct the 3D spectral cube from the SCI measurement and masks.
- The self-attention generative adversarial network (GAN) [40] is integrated with the U-net to exploit the non-local correlation in the hyperspectral images. This self-attention GAN plus U-net constitutes the *reconstruction stage* of our network.
- A *refinement stage* composed of a small U-net and residual learning [41] is developed to boost up the quality of reconstructed images from the first stage. In this stage, each channel is performed independently.
- We have verified our proposed network on extensive “real-mask-in-the-loop" simulation data and also the *real data* captured by our camera [30]. Our network offers much better results than DeSCI for simulation data and higher performance for real data, and it finishes the reconstruction in sub-seconds while DeSCI needs hours [42].

CHAPTER 2

Direct Shape Regression Networks for End-to-End Facial Landmarks Detection

Firstly, we talk about the related work. Then, we introduce our direct shape regression network (DSRN). We start with the problem formulation (2.2) and describe in detail the key components of DSRN, that is, the doubly convolutional layer (2.3), the Fourier embedding layer (2.4) and the low-rank learning layer (2.5). Finally, we will show the experimental results and discuss our DSRN.

2.1 Related Work

facial landmarks detection has been extensively studied and remarkable progress has occurred over the past decades [43, 44, 45, 46]. Previous work mainly focused on cascaded regression, which relies on iterative optimization. Cascaded regression starts with an initial shape which can be a random guess or the mean shape of training samples, and iteratively refines the shape by a cascade of regressors. Building upon cascaded regression, many improved variants have been developed which distinguish themselves by the shape initialization strategies [46], shape-indexed features [47] or regressors [48].

Xiong *et al.* proposed a supervised descent method (SDM) [48] to address the cascaded regression problem by optimizing non-linear least squares based on SIFT [24] features. Zhu *et al.* use a coarse-to-fine shape searching method to locate the landmarks. That method is robust to large pose variation [49]. To achieve high performance, they employ multiple hybrid handcrafted features, e.g., SIFT, HOG and BRIEF etc, as local descriptors. Support vector regression and random forests are used by [50] for facial landmarks

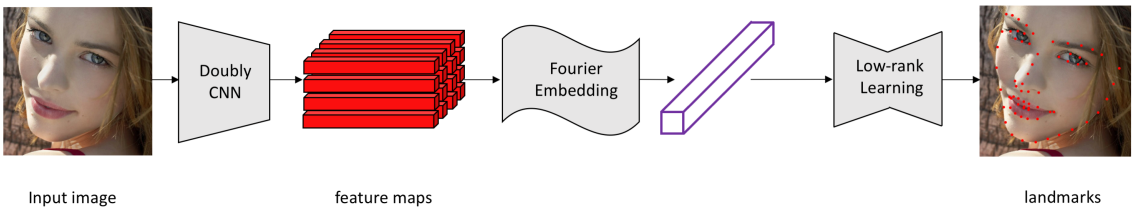


Figure 2.1: The architecture of our proposed direct shape regression network).

detection from the local image patch. By using Markov random field to model the spatial relations of landmarks, they try to resolve the predictions uncertainties.

With the great success of deep learning in feature representation, some methods use convolutional neural networks (CNNs) to learn the features or deep models to represent the regressors. Sun *et al.* [51] constructed a deep convolutional network cascaded structure to detect facial points, with multi-level regression networks. Liu et al.*et al.* [52] not only consider the spatial domain, but also use recurrent neural networks (RNN) to get the temporal information in the video-based facial landmarks detection datasets.

However, most deep learning-based models are still based on cascaded regression, which is sensitive to improper shape initialization. Some recent methods [46, 53] attempt to solve this problem by running algorithms more than one times, but the dependence on shape initialization is still not totally avoided. Lv *et al.* [54] use a two-stage regression method. It uses spatial transformer networks [55] to transform the full face and face parts to canonical shape respectively in two stages. They call this step re-initialization. However, this method does not optimize the network parameters in the two stages jointly. The first end-to-end recurrent convolutional system for facial landmarks detection was proposed in [56]. They use CNNs to extract features and a connected RNN to approximate the cascaded process. The main difference from our end-to-end learning is that our method is direct shape regression which starts with a raw image and directly predicts coordinates of landmarks on facial shapes rather than estimating shape increments iteratively. Bulat *et*

al. [57] propose a method that can also map 2D facial landmarks to 3D. We should also mention the method of [58], which is a facial alignment method explicitly designed to be lightweight and suitable for devices with limited computational resources. Obviously, our method has a different scope as it is designed for usage with modern desktop computers.

In contrast to those existing methods, our DSRN is, to the best of our knowledge, the first method that achieves direct shape regression in an end-to-end learning framework, without relying on cascaded regression. DSRN addresses the central issue of face landmarks detection by effectively disentangling the highly nonlinear relationship between images and facial shapes while simultaneously encoding correlations of landmarks on the shape. It leverages the strengths of neural networks for structured prediction and kernels for nonlinear feature extraction.

2.2 Preliminaries

Facial landmarks detection is the task of finding a mapping from an input image I to the facial shape S represented by the coordinates of landmarks in the form of a vector, $[x_1, y_1, ..., x_N, y_N]^\top \in \mathbb{R}^{2N}$, where N is the number of landmarks. DSRN directly predicts shapes from images in an end-to-end learning architecture, which handles major challenges of facial landmarks detection in one single framework. Specifically, the doubly convolutional layer in conjunction with the Fourier embedding layer are used for effective nonlinear feature extraction, to model the nonlinear relationship between images and shapes; the linear low-rank learning layer explicitly encodes intrinsic correlations of landmarks in a data-driven way for robust and improved estimation.

2.3 Doubly Convolutional Layer

Image representation plays a fundamental role in facial landmarks detection. Hand-crafted features, e.g., SIFT [24] and HoGs [25], were extensively used in previous methods [49, 59]. The convolutional neural network (CNN) has recently emerged as a powerful tool for feature extraction and shown great success in diverse visual tasks.

However, the size of training data is relatively small in facial landmarks detection, while images exhibit great appearance variation and face shapes show huge variability. This poses great challenges to conventional CNNs. Instead of using regular convolutions, we use a doubly convolutional module [26], which has shown improved performance in term of both efficiency and effectiveness. The double convolution is inspired by the fact that many of filters in regular convolutions are very similar or almost translated version of each other, which induces huge redundancy. It can largely reduce the number of parameters while improving the performance, which is well suited for facial landmarks detection.

In double convolutions, there are a set of meta filters with size $L^{'} \times L^{'}$. The size of effective filters is $L \times L$ where $L < L^{'}$. So we can consider that there are $(L^{'} - L + 1)^2$ effective filters within each meta filter, and the group of effective filters are forced to be translated versions of each other. When the input image is convolved with one meta filter, it convolves with each effective filter in this meta filter, to produce $(L^{'} - L + 1)^2$ feature maps for this meta filter. As a consequence, we use only one meta filter with $L^{'} \times L^{'}$ parameters, while obtaining the same number of feature maps as using $(L^{'} - L + 1)^2$ individual filters with $(L^{'} - L + 1)^2 \times L \times L$ parameters. The structure of doubly convolutional layer is shown in Fig 2.2.

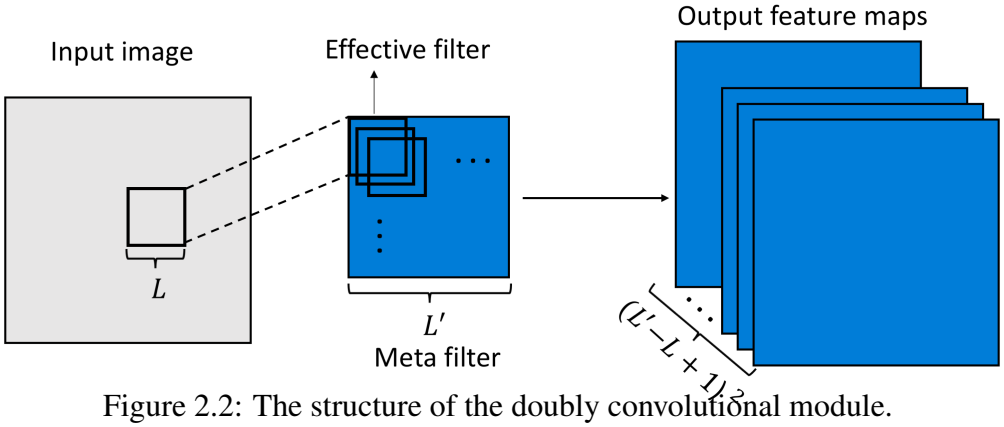


Figure 2.2: The structure of the doubly convolutional module.

2.4 Fourier Embedding Layer

To handle the complicated relationship between images and facial shapes, nonlinear feature extraction is usually required to achieve high-level representations. The doubly convolutional module produces a set of feature maps contained in $X \in \mathbb{R}^{w \times h \times c}$ with width w, height h and the number of maps c. For a c dimensional vector of a spatial location across the feature maps in X, we use notation $\mathbf{x} = [x_1, x_2, ..., x_c]^\top \in \mathbb{R}^c$. We need to pool those $w \times h$ c-dimensional feature vectors into in a holistic representation for shape regression.

In this work, we propose using Fourier embedding to pool feature maps by leveraging the great strength of kernels for nonlinear feature extraction, which enables filling the semantic gap between images and shapes. The Fourier embedding is derived from the approximation of shift invariant kernels [60, 61] which is underpinned by Bochner's Theorem [62].

Theorem 1 (Bochner [62])**.** *A continuous shift-invariant kernel function* $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$ *on* $\mathbb{R}^d$ *is positive definite if and only if it is the Fourier transform of a unique finite non-negative measure on* $\mathbb{R}^d$*. Defining* $\zeta_\omega(\mathbf{x}) = e^{j\boldsymbol{\omega}^\top \mathbf{x}}$*, for any* $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$*,*

$$k(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^d} p(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')} d\boldsymbol{\omega} = \mathbb{E}_{\boldsymbol{\omega}}[\zeta_{\boldsymbol{\omega}}(\mathbf{x}) \zeta_{\boldsymbol{\omega}}(\mathbf{x}')^*] \tag{2.1}$$

*where * is the conjugate and $p(\boldsymbol{\omega})$ is the Fourier transform of the kernel.*

The kernel $k(\mathbf{x}, \mathbf{x}')$ can be approximated by drawing d random samples,

$$k(\mathbf{x}, \mathbf{x}') \approx \sum_{i=1}^{d} \left\langle \sqrt{\frac{2}{d}} \cos(\boldsymbol{\omega}_i^\top \mathbf{x} + b_i), \sqrt{\frac{2}{d}} \cos(\boldsymbol{\omega}_i^\top \mathbf{x}' + b_i) \right\rangle \tag{2.2}$$

where $\boldsymbol{\omega}$ is sampled from the probablity distribution $p(\boldsymbol{\omega})$, and b is uniformly sampled over $[0, 2\pi]$.

Therefore, the corresponding approximated feature map $\boldsymbol{\phi}(\mathbf{x})$ is

$$\boldsymbol{\phi}(\mathbf{x}_i) = \sqrt{\frac{2}{d}} [\cos(\boldsymbol{\omega}_i^\top \mathbf{x}_i + b_i)]_{1:d} \tag{2.3}$$

where $\boldsymbol{\phi}(\mathbf{x})$ is called the random Fourier feature [60], and has been successfully used in various kernel methods.

However, the great power of kernel approximation based on random Fourier features remains largely underdeveloped, and this topic has recently attracted attention [63]. In most of the existing kernel approximation methods, the sampling is independent of input data distributions, and this usually requires high-dimensional feature maps to achieve kernel approximation with satisfactory performance. Moreover, since no learning is involved, the approximate feature maps would be of high redundancy and of low discriminant ability, which compromises performance while inducing unnecessary computational cost. In addition, approximating the kernel with a fixed configuration does not necessarily lead to high performance since it remains an open question how to choose the best kernel configuration.

Instead of approximating kernels by random sampling from data-independent distributions, we learn the parameters $\{\boldsymbol{\omega}, b\}$ from data in a supervised way, which enables more compact but highly discriminative feature representations.

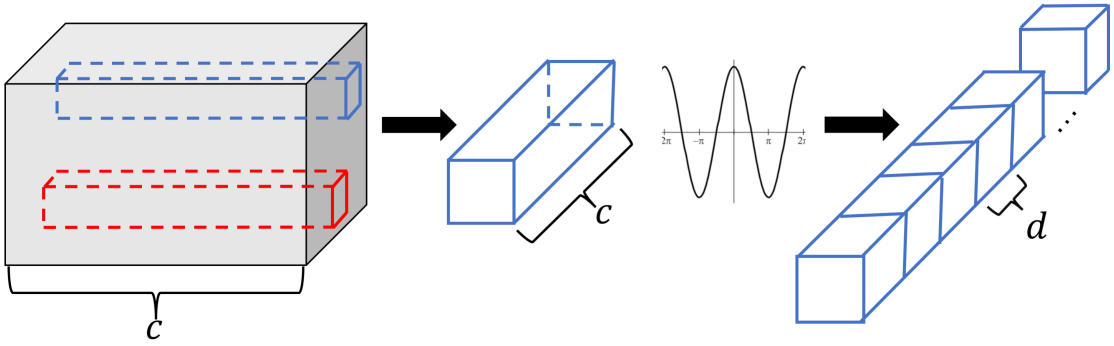


Figure 2.3: The structure of Fourier embedding.

Define $W = [\boldsymbol{\omega}_1, \cdots, \boldsymbol{\omega}_d] \in \mathbb{R}^{d \times c}$ and $\boldsymbol{b} = [b_1, \cdots, b_d]$. We define a nonlinear layer of neural networks with cosine activations,

$$\boldsymbol{\phi}_i = \cos(W\mathbf{x}_i + \mathbf{b}) \tag{2.4}$$

where $\cos$ is an element-wise function, i indicates the i-th location in the feature map X and W is the weight matrix of the nonlinear layer. The induced Fourier embedding layer can be seamlessly integrated with the doubly convolutional layer to achieve a fully end-to-end learning architecture that can be trained via back-propagation.

To achieve a holistic representation, we concatenate the embedded feature vectors into a single vector $\mathbf{z} = [\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_i, \cdots, \boldsymbol{\phi}_p] \in \mathbb{R}^D$, where $p = w \times h$, i.e., the number of locations. In contrast to feature pooling techniques by directly summing up the feature vectors, the concatenation can well preserve the spatial information of images, which is of great importance for predicting the spatial locations of facial landmarks.

2.5 Low-rank Learning Layer

We propose a simple but effective linear layer to encode correlations of landmarks by low-rank learning. Having the holistic representation $\mathbf{z}$, a straightforward way for prediction is to use a fully connected layer with the regression matrix represented by $M \in \mathbb{R}^{Q \times D}$,

where Q is the number of outputs, i.e., $Q = 2N$, which gives $\mathbf{y} = M\mathbf{z}$. An identity activation function is used by default. Although sharing the holistic representations, landmark correlations are not explicitly encoded. Low-rank constraints, such as the nuclear norm, could be simply imposed to force the regression matrix M to be low rank, but this does not always guarantee low-rankness of M, and can fail to fully capture the correlations. Instead of using one fully connected layer, we propose linear low-rank learning layer to explicitly encoding correlations of landmarks.

Specifically, we propose the low-rank learning layer by replacing the single matrix M with multiplication of two low-rank matrices, which gives rise to

$$\mathbf{y} = M\mathbf{z} = U^\top V\mathbf{z} \tag{2.5}$$

where $U \in \mathbb{R}^{P\times Q}$, $V \in \mathbb{R}^{P\times D}$ and $P \leq Q$. The linear function provides a low-rank layer to explicitly encode inter-output correlations. U and V are learned in a data-driven way without relying on any specific assumptions, and can adaptively capture specific correlations in different applications.

Low-rank learning brings two attractive advantages compared to nuclear norm based minimization. First, it establishes an overall mapping of M with guaranteed low rankness to explicitly encode correlations; related outputs are forced to share similar regression parameter patterns, and thus knowledge is transferred across correlated outputs. This can significantly improve the overall prediction performance. Second, low-rank learning avoids solving complicated rank-constrained problems and leverages the great effectiveness of linear learning, which enjoys great computational efficiency; by setting $P \ll Q$, the low-rank learning can greatly reduce the number of parameters, which is especially advantageous when using iterative optimization with stochastic gradient descent [64].

2.6 Experiments and Results

We have conducted extensive experiments on five benchmark datasets, and we provide a comprehensive comparison with state-of-the-art methods. The proposed direct shape regression network (DSRN) consistently yields high accuracy for facial landmarks detection, and in most cases outperforms previous methods by large margins. Moreover, the consistently high performance on the five diverse facial landmarks detection tasks demonstrates the generality of our method.

2.6.1 Datasets

The five datasets are commonly used benchmarks for facial landmarks detection. Faces in the datasets are collected in uncontrolled scenarios, demonstrating great variations, which pose significant challenges for facial landmarks detection. We provide the detailed description of those datasets to facilitate direct comparison with previous work under the same experimental settings.

AFLW [65] contains a total of 24386 face images gathered from Flickr. In contrast to other databases limited to frontal views or acquired under controlled conditions. AFLW faces are collected in the wild, have large-scale pose variations up to $\pm 90^{\circ}$ and also have large variety in face appearance (e.g., pose, expression, ethnicity, gender). Each image is annotated with 21 landmarks. Following the experimental settings of cascaded compositional learning (CLL) [47], we ignore the two ear points and use the same 20000 and 4386 images for training and test, respectively.

300W [66, 67] consists of several datasets including AFW [45], HELEN [68], LFPW [23], XM2VTS [69]. In addition, it contains a challenging 135-image IBUG [70] set. Following the same dataset configuration in [49], our training set of 3148 images consists of the full set of AFW and the training sets of HELEN and LFPW. The full test set (689 images) is divided into a "common subset" (554 images), which contains the test sets from LFPW

Method	Error	Year
CDM [73]	5.43	2013
PCPR [53]	3.73	2013
ERT [74]	4.35	2014
SDM [48]	4.05	2013
LBF [75]	4.25	2014
PO-CR [76]	5.32	2015
CFSS [49]	3.92	2015
CLL [47]	2.72	2016
DAC-CSR [59]	2.27	2017
DRA-TSR [54]	2.17	2017
DSRN	**1.86**	

Table 2.1: Comparison on AFLW.

and HELEN, and a "challenging subset" (135 images) which is from IBUG. 300W has a 68-points annotation for each face image.

CelebA [71] is a large-scale face dataset with 202599 images. CelebA provides 5 five landmarks of the facial shape for each image. The images show large pose variations and background clutter. Because of large diversities and large quantities, CelebA is suitable for training and testing a deep learning model. Following the original work [71], 182631 and 19926 images are used respectively for the training and test sets.

MAFL is a subset of CelebA. In order to benchmark with previous methods, we following the experimental settings in [72]. Specifically, we sample the same 20000 faces from CelebA and select the same 1000 faces for testing as in [72].

300VW [70] is a video-based facial landmarks detection dataset which contains 114 videos from different conditions. We extract face images from the same 50 videos as [70] to train the model, and the remaining 64 videos are divided into three test sets.

2.6.2 Implementation Details

We use four doubly convolutional layers and four pooling layers for the feature extraction task. Multiple feature maps are produced in each convolutional layer. Follow-

ing each convolution operation, we use rectified linear unit as activation function and the $5 \times 5, 5 \times 5, 3 \times 3, 3 \times 3$ max pooling. After that, the Fourier embedding layer is added to the feature maps $X \in \mathbb{R}^{8\times8\times256}$. In Fourier embedding, we get $X^{'} \in \mathbb{R}^{8\times8\times d}$ fisrt, where the value of d may be changed depending on the size of training samples and the number of landmarks in the task. Then we do concatenation for $X^{'}$.

In the low-rank learning layer, we do not use any nonlinear activtion functions but just the linear function with identity activations. The commonly used weight decay and batch normalization [77] techniques are also used. The parameter for weight dacay is 0.001. We employ the stochastic optimization algorithm Adam [78] to learn the parameters of the neural network. The learning rate starts from 0.0005 and with 0.95 exponential decay every 10000 iterations, and the mini-batch size is set to 64.

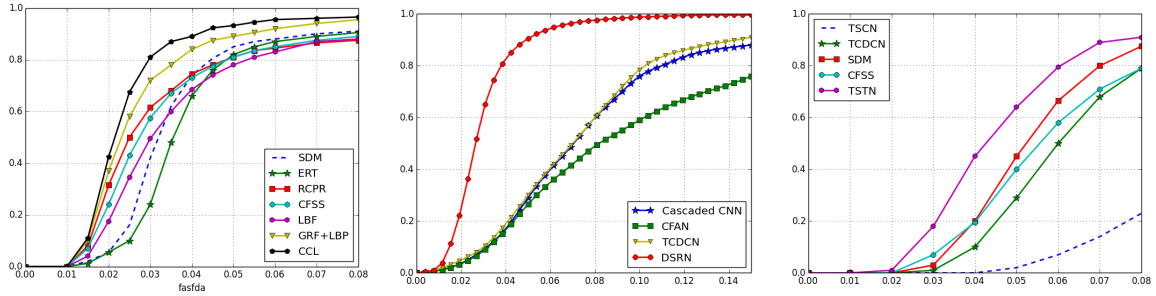
For all experiments, the original bounding box given by the dataset is used, without any data augmentation. For the 300W dataset, due to the size of the training set being relative small, we pre-train our model on the large-scale 300VW dataset which has the same number, 68, of landmarks, and fine tune it on the training set of 300W to obtain the final model.

We use the normalized mean error (NME) as the evaluation metric, which is defined as follows:

$$\mathrm{NME} = \frac{\frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}}{d}, \tag{2.6}$$

where (x, y) and $(\hat{x}, \hat{y})$ denotes the ground truth and predicted coordinates, respectively, N denotes the number of landmarks on facial shapes, and d is the distance for normalization.

Following previous work, for 300W, CelebA, MAFL and 300VW, we use the inter-ocular distance to normalize the mean error; for AFLW, we use face size to normalize mean error since the inter-ocular distance of many faces is close to zero. For brevity, $\%$ is omitted



AFLW MAFL 300VW

Figure 2.4: Comparions on AFLW, MAFL and 300VW in terms of CED.

Method	Common Subset	Challenging Subset	Full Test set
RCPR [53]	6.18	17.26	8.35
SDM [48]	5.57	15.40	7.52
ESR [46]	5.28	17.00	7.58
GN-DPM [79]	5.78	-	-
ERT [74]	-	-	6.40
CFAN [80]	5.50	16.78	7.69
LBF [81]	4.95	11.98	6.32
DDN [82]	-	-	5.59
CFSS [49]	4.73	9.98	5.76
MDM [56]	4.83	10.14	5.88
DRA-TSR [54]	4.36	**7.56**	**4.99**
DSRN	**4.12**	9.68	5.21

Table 2.2: Comparison on 300W.

in all tables. We also show the evaluation results in the form of cumulative error distribution (CED) curve for comprehensive comparison.

2.6.3 Performance and Comparison

Our DSRN consistently achieves high performance on the five datasets and surpasses previous methods in most cases.

On AFLW, as shown in Table 2.1, DSRN achieves the best error rate, 1.86%, compared to the previous best error rate of 2.17% [54]. In Fig 2.4 (a), the curve of our DSRN

Method	CelebA	MAFL
TCDCN [72]	-	7.95
Cascaded CNN [51]	-	9.73
CFAN [80]	-	15.84
RCPR [53]	4.12	-
SDM [48]	4.35	-
CFSS [49]	3.95	-
DSRN	**3.08**	**3.15**

Table 2.3: Comparison on CelebA and MAFL.

Method	Test 1	Test 2	Test 3	Year
SDM [48]	7.41	6.18	13.04	2013
TSCN [83]	12.54	7.25	13.13	2014
CFSS [49]	7.68	6.42	13.67	2015
TCDCN [72]	7.66	6.77	14.98	2016
TSTN [52]	5.36	**4.51**	12.84	2017
DSRN	**5.33**	4.92	**8.85**	-

Table 2.4: Comparison on 300VW.

is clearly above those of other methods, which also indicates the performance advantages. Compared with those methods based on cascaded regression, our DSRN can detect the landmarks for side faces accurately as shown by the intuitive illustration in the fourth and seventh images of Fig 2.5 (a).

On 300W, our DSRN achieves competitive performance, which is better than all previous methods except for [54], which gives better results on the challenging set and the full test set. The challenges of 300W stem from the great variations of images while with limited training data. As shown in Fig 2.5, our DSRN can accurately predict the landmarks on faces with large orientations and diverse expressions.

On CelebA and MAFL, as can be seen in Table 2.3, our DSRN achieves the best performance on both datasets, with error rates of 3.08% and 3.15% respectively, which are significant improvements over the previous best error rates of 3.95% and 7.95% respectively. In Fig 2.4 (b), we can see that there is a big gap between DSRN and TCDCN, which uses the similar convolutional network with DSRN and takes advantage of face attributes,

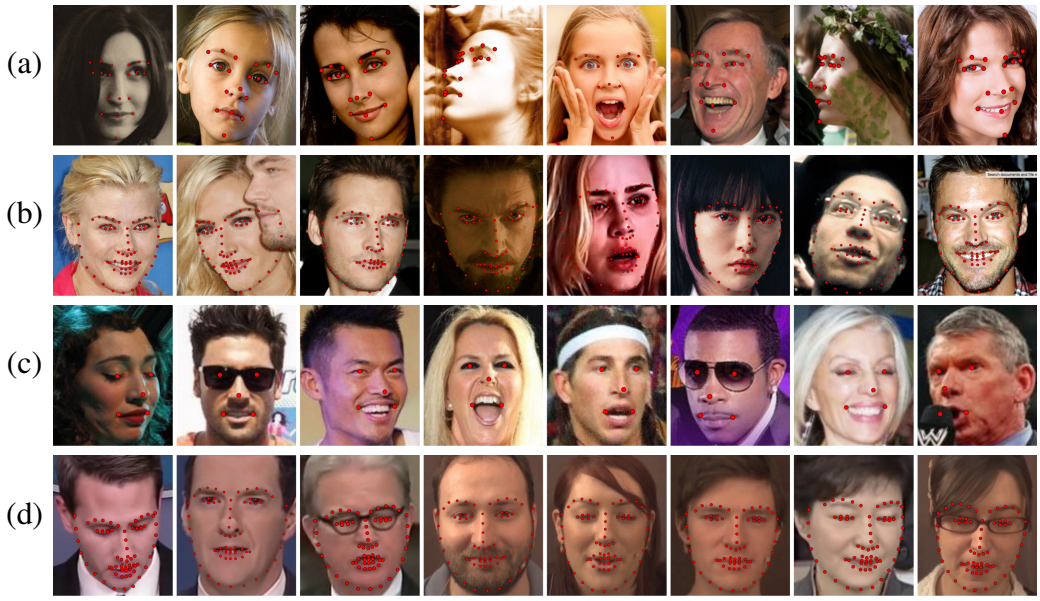


Figure 2.5: Illustrative results on (a) AFLW (b) 300W (c) CelebA (d)300VW .

but without Fourier embedding and low-rank learning layer. In the second and sixth images of Fig 2.5 (c), when the eyes in face images are occluded by sunglasses, DSRN can still predict the landmarks correctly. This is because our low-rank learning can encode the intrinsic correlation of landmarks.

On 300VW, as shown in Table 2.5, DSRN produces the highest accuracy on Test 1 and 3, where Test 3 is regarded as the most challenging subset. We have also compared with TSTN [52] designed specifically for video-based facial landmarks detection by modeling the temporal relationship across frames. Our method achieves overall better performance than TSTN. Moreover, DSRN can run very fast with about 500 frames per second excluding face detection in the platform of NVIDIA GTX 1080Ti GPU, which is promising for the prospect of practical application. The intuitive results of 300VW are shown in Fig 2.5, our DSRN can accurately predict the shapes of face images with great appearance variations.

Since the code for most methods are not realse, we only compare the speed for our network with MDM and DRA-TSR, the result is showed in table 2.6. Titan X and 1080Ti

	Our network	MDM	DRA-TSR
speed (fps)	500	270	110
platform	1080 Ti	1080Ti	Titan X
year	2018	2017	2017

Table 2.5: Comparison of speed for different methods.

have the same computational ability and the major difference is the memory. So we can conclude that our network is much faster than the state of art methods.

2.7 Discussion

In this part, we propose the direct shape regression network (DSRN) for end-to-end facial landmarks detection. DSRN consists of the doubly convolutional layer, the novel Fourier embedding layer, and the low-rank learning layer. These layers enable jointly handling nonlinear image-shape relationships and the intrinsic correlations between landmarks. Our DSRN offers a new learning architecture that combines the strengths of kernels for nonlinear feature extraction and neural networks for structured prediction. Experimental results on five benchmark datasets have shown that our DSRN delivers high performance on all datasets. The effectiveness of DSRN on the diverse facial landmarks detection tasks and the fast inference time offer promise for real time landmarks prediction task.

CHAPTER 3

Images Reconstruction for Compressive Imaging System

For the snapshot-spectral compressive imagaing system like CASSI system [28, 29], the spectral scene is collected by the objective lens and spatially coded by a fixed mask. Then the coded scene is spectrally dispersed by the disperser. Following this, the spatial-spectral coded scene is detected by the charge-coupled device (CCD). A snapshot on the CCD thus encodes tens of spectral bands of the scene. The number of coded frames for a snapshot is determined by the dispersion property of the dispersive element and the pixel sizes of the mask and the CCD.

For the snapshot-video compressive imagaing system like CACTI system [84], the objective lens will collect the high-speed scene. Then it will be spatially coded by the mask. The monochrome or color CCD will detect the coded scene for grapyscale and color video capturing. Tens of temporal frames of the high-speed scene will be encoded by the snapshot on the CCD.

Consider B-frames are modulated and encoded in SCI and each frame has n ($= n_x \times n_y$) pixels. Without considering optical details, mathematically, the measurement in SCI can be modeled by [29]

$$\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{g}\,, \tag{3.1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{n \times nB}$ is the sensing matrix, $\boldsymbol{x} \in \mathbb{R}^{nB}$ is the desired signal, and $\boldsymbol{g} \in \mathbb{R}^{n}$ denotes the noise.

Though Eq. (3.1) has the formulation similar to compressive sensing (CS) [85, 86], unlike traditional CS, the sensing matrix considered here is not a dense matrix, and it does

not satisfy the Restricted Isometry Property *(RIP)*. In SCI, the matrix $\mathbf{\Phi}$ has a very specific structure and can be written as

$$\mathbf{\Phi} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_B], \tag{3.2}$$

where $\{\boldsymbol{D}_k\}_{k=1}^{B}$ are diagonal matrices defined by the following mask. Specifically, consider that B spectral frames $\{\boldsymbol{X}_k\}_{k=1}^{B} \in \mathbb{R}^{n_x \times n_y}$ are modulated by shifted versions of the fixed mask, $\{\boldsymbol{C}_k\}_{k=1}^{B} \in \mathbb{R}^{n_x \times n_y}$, correspondingly. The measurement $\boldsymbol{Y} \in \mathbb{R}^{n_x \times n_y}$ is given by

$$\boldsymbol{Y} = \sum_{k=1}^{B} \boldsymbol{X}_k \odot \boldsymbol{C}_k + \boldsymbol{G}, \tag{3.3}$$

where $\odot$ denotes the Hadamard (element-wise) product, and $\boldsymbol{D}_k = \mathrm{diag}(\mathrm{vec}(\boldsymbol{C}_k))$, for $k = 1, \ldots, B$. For all B pixels (in the B frames) at position (i, j), $i = 1, \ldots, n_x$; $j = 1, \ldots, n_y$, they are collapsed to form one pixel in the snapshot measurement as $y_{i,j} = \sum_{k=1}^{B} c_{i,j,k} x_{i,j,k} + g_{i,j}$. By defining $\boldsymbol{x} = \left[\boldsymbol{x}_1^{\mathsf{T}}, \ldots, \boldsymbol{x}_B^{\mathsf{T}}\right]^{\mathsf{T}}$, where $\boldsymbol{x}_k = \mathrm{vec}(\boldsymbol{X}_k)$, we have the vector formulation of Eq. (3.1). Thus, $\boldsymbol{x} \in \mathbb{R}^{n_x n_y B}$, $\mathbf{\Phi} \in \mathbb{R}^{n_x n_y \times (n_x n_y B)}$, and the compressive sampling rate in SCI is equal to $1/B$. It has been proved recently in [87] that the reconstruction of SCI is bounded even when $B > 1$.

3.1 Reconstruct Hyperspectral Images from a Snapshot Measurement

The target of network is to reconstruct the hyperspectral image cube from the single measurement captured by the SCI camera. Recently, GAN [88] and variational autoencoder (VAE) [89] become the most convening generative models. Recent researches have suggested that using U-net as the generative model in GAN is capable of solving diverse problems [90, 91, 92]. In our task, in addition to the U-net plus GAN, the most recently proposed self-attention mechanism is adapted to exploit both the non-local similarity of spatial textures and the long-range spectral similarity. We propose an additional HCR strategy to gradually reconstruct all channels which guarantees the quality of result and the accuracy of spectral information.

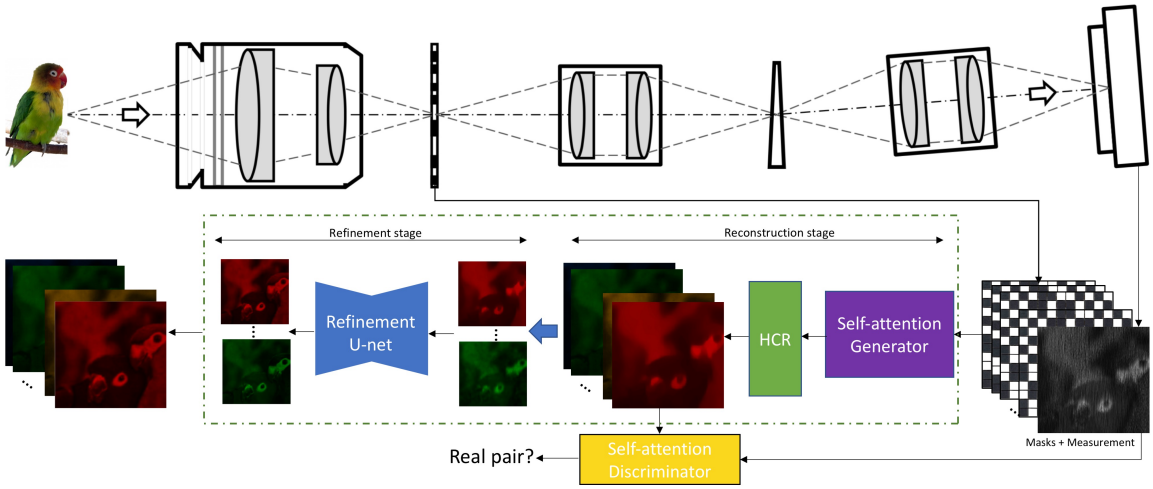


Figure 3.1: Imaging process of SCI and the network structure

Network reconstructs hyperspectral images with B (=24 in our experiments) spectral channels that is high dimensional data. Even a deep U-net and HCR are used, it still does not guarantee to reconstruct high quality images due to the large number of parameters and the limited training data. In order to overcome this challenge, we propose to use another *refinement* U-net which is shallower than the first U-net in the reconstruction stage. This refinement stage improves the image quality of each spectral channel separately.

3.1.1 Reconstruction Stage

The reconstruction stage outputs the hyperspectral images and it aims to extract both spatial and spectral information from the measurement.

3.1.1.1 Conditional GAN

The discriminator in conditional GAN (cGANs) [93] can also observe the inputs from the generator. cGAN is appropriate for our SCI reconstruction as we aim to generate corresponding output hyperspectral images conditional on the input measurement and masks. Specifically, the inputs masks are fixed (in a pre-built SCI system) while the input

measurement depends on the captured scene. Thereby, the masks are not necessary to be observed by the discriminator. The objective function of our cGAN can be expressed as

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}}(G, D) = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}[\log D(\boldsymbol{y}, \boldsymbol{x})] \\ + \mathbb{E}_{\boldsymbol{y}}[\log(1 - D(\boldsymbol{y}, G(\boldsymbol{y}, \boldsymbol{\Phi})))],\end{aligned} \tag{3.4}$$

where G and D denotes the generator and discriminator, respectively.

3.1.1.2 Deeper U-net with Self-Attention

The U-net architecture detailed in Fig. 3.2 is used as the generator in our cGAN. As mentioned before, our output hyperspectral images and the input measurement share the same spatial structure, *e.g.*, location of edges. The encoder and decoder can help capture the shared low level information between input and output and remove the noise; but it may also lose the location information from the measurement. To tackle this challenge, we add the skip connection to help the location information pass through the network. Furthermore, since we are reconstructing high dimensional hyperspectral images, we employed a deeper U-net. In particular, we have 3 times convolution operations with stride 1 after the downsampling or upsampling (which is 2 in [39]); we also have 5 times downsampling and upsampling in the encoder and decoder of U-net instead of 4. Experiment results in Sec. 3.2.3.1 (Table 3.2) show that our deeper U-net achieves better (1.92dB higher PSNR) results than the original U-net.

Attention module has been widely used in many computer vision tasks. Since the convolution operator in U-net has a local receptive field, only multiple convolutional layers can capture the long range dependencies. Via adding the self-attention, the network can learn the long range similarity in one layer easily. Self-attention learns an attention map [94] which represents the extent that depends on all other location pixels when generating a specific location. In our self-attention layer, all spectral channels share the same

attention map, as we not only want to capture the long range dependencies in space but also to keep the spectral similarity in SCI reconstruction. The self-attention (Fig. 3.3) is not only used in the generator but also in the discriminator. We have performed the experiments by adding self-attentions to different layers of the network and found that imposing it on the middle-to-high layer feature maps will lead to better results, but with larger attention maps. Limited by the GPU memory, we show results by imposing the self-attention to the layer who has 256 feature maps before the deconvolution in the decoder of the U-net in Fig.3.

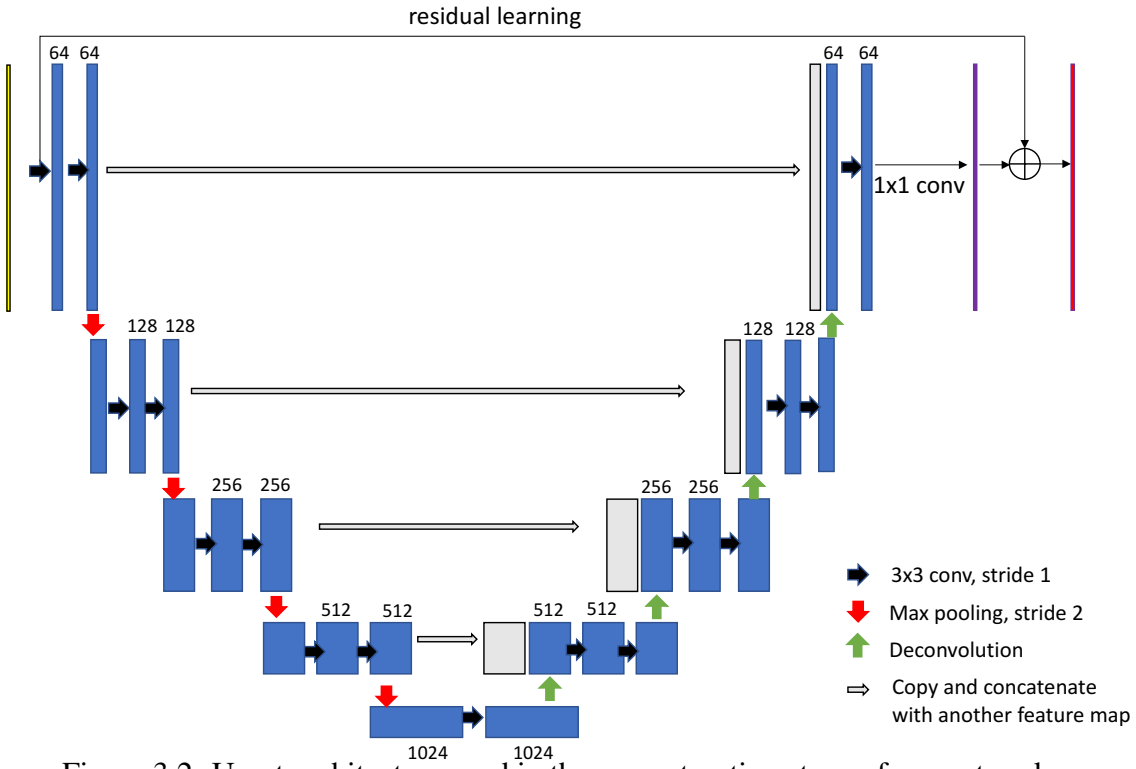


Figure 3.2: U-net architecture used in the reconstruction stage of our network.

As depicted in Fig. 3.3, let $\boldsymbol{\theta} \in \mathbb{R}^{c\times h\times w}$ denote the feature map that we want to impose the self-attention. By using 1×1 convolutions on $\boldsymbol{\theta}$, we can get three feature spaces

$$f(\boldsymbol{\theta}) \in \mathbb{R}^{c'\times h\times w},\ \ g(\boldsymbol{\theta}) \in \mathbb{R}^{c'\times h\times w},\ h(\boldsymbol{\theta}) \in \mathbb{R}^{c\times h\times w},$$

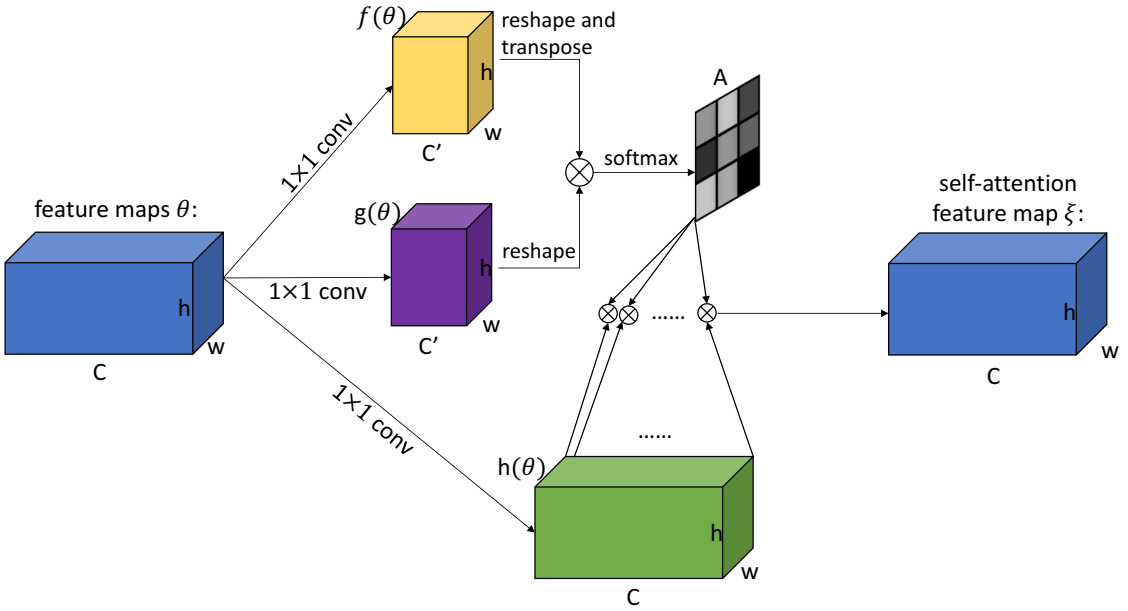


Figure 3.3: The self-attention module in our framework.

where c' is an integer and we set $c' = \frac{c}{8}$ in our experiments. We now use $\{f(\boldsymbol{\theta}),\ g(\boldsymbol{\theta})\}$ to calculate the attention map. First, we reshape them to 2D matrices $\{f'(\boldsymbol{\theta}),\ g'(\boldsymbol{\theta})\} \in \mathbb{R}^{c' \times N}$, with $N = h \times w$; then each entry of the attention map $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is calculated by

$$a_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})},\ \text{with}\ \ s_{ij} = f'(\theta_i)^{\mathsf{T}} g(\theta_j). \tag{3.5}$$

$a_{j,i}$ represents that the extent of the model depends on the i^{th} location when generating the j^{th} region. This attention map $\boldsymbol{A}$ is then incorporated with the feature space $h(\boldsymbol{\theta})$. We first reshape $h(\boldsymbol{\theta})$ to $h'(\boldsymbol{\theta}) \in \mathbb{R}^{c \times N}$ and then impose $\boldsymbol{A}$ on it, which arrives

$$\boldsymbol{\xi}' = \boldsymbol{A}\ h'(\boldsymbol{\theta})^{\mathsf{T}} \in \mathbb{R}^{N \times c}. \tag{3.6}$$

Following this, we reshape each channel (column) in $\boldsymbol{\xi}'$ to get the output of the attention layer $\boldsymbol{\xi} \in \mathbb{R}^{c \times h \times w}$. Lastly, then, we multiply the output of the attention layer $\boldsymbol{\xi}$ by a *scale learnable* parameter γ and add it back to the input feature map $\boldsymbol{\theta}$. This leads to the final result

$$\boldsymbol{z} = \gamma \boldsymbol{\xi} + \boldsymbol{\theta}. \tag{3.7}$$

3.1.1.3 Hierarchical Channel Reconstruction

It is challenging to reconstruct all 24 channels images from a single measurement in one shot. Therefore, we propose a progressive reconstruction scheme, *i.e.*, Hierarchical Channel Reconstruction (HCR). HCR tries to recover a fraction of the spectral channels and then reconstruct the entire channels based on the information we have recovered.

In our experiment, 24 spectral channels need to be reconstructed. We first reconstruct $[\boldsymbol{x}_1, \boldsymbol{x}_5, \boldsymbol{x}_9, \boldsymbol{x}_{13}, \boldsymbol{x}_{17}, \boldsymbol{x}_{21}]$ spectral channels with an interval of 4. Then we reconstruct the $[\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5, \ldots, \boldsymbol{x}_{23}]$ spectral channels with an interval of 2. Finally, all the 24 channels are reconstructed. The residual learning method is also employed. Details of the proposed HCR are showed in Fig. 3.4. In this manner, our network reconstructs the hyperspectral images gradually, where we have decomposed the $1 \rightarrow 24$ problem to $1 \rightarrow 6 \rightarrow 12 \rightarrow 24$ cascaded problems. In other words, if we can reconstruct partial spectral channels with correct spectral information, a simple interpolation method should be qualified to reconstruct the entire channels. Table 3.2 shows HCR has improved the performance of λ -net.

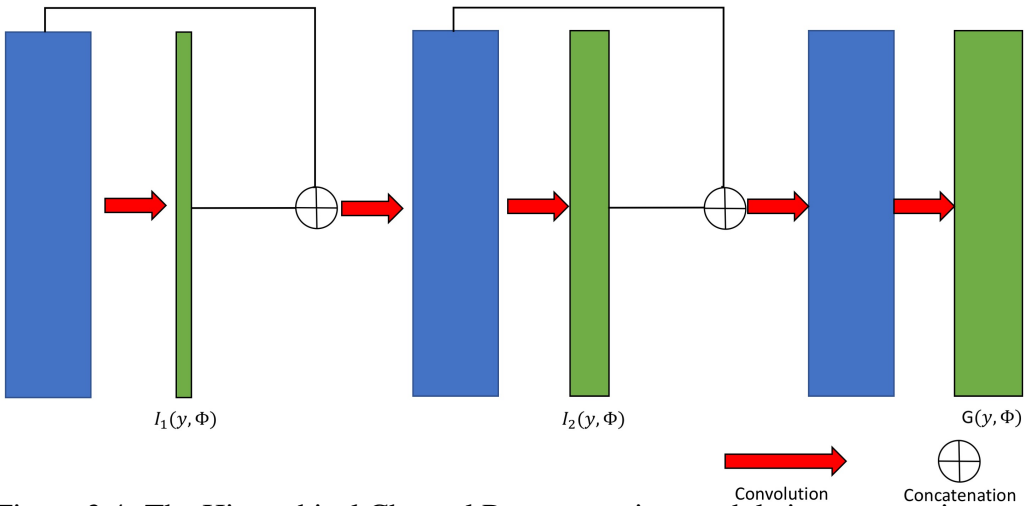


Figure 3.4: The Hierarchical Channel Reconstruction module in our experiment.

We define the intermediate outputs and the final output as $I_1(\boldsymbol{y}, \boldsymbol{\Phi}), I_2(\boldsymbol{y}, \boldsymbol{\Phi})$ and $G(\boldsymbol{y}, \boldsymbol{\Phi})$, respectively. The target of network is to reconstruct the signal and thus it is reasonable to add the ℓ_2 loss into our objective function,

$$\begin{aligned}\mathcal{L}_{\ell_2}(G) = \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}}[\|\boldsymbol{x}^1 - I_1(\boldsymbol{y}, \boldsymbol{\Phi})\|_2 + \|\boldsymbol{x}^2 - I_2(\boldsymbol{y}, \boldsymbol{\Phi})\|_2 \\ + \|\boldsymbol{x} - G(\boldsymbol{y}, \boldsymbol{\Phi})\|_2],\end{aligned} \tag{3.8}$$

where $\boldsymbol{x}^1 \overset{\text{def}}{=} [\boldsymbol{x}_1, \boldsymbol{x}_5, \boldsymbol{x}_9, \boldsymbol{x}_{13}, \boldsymbol{x}_{17}, \boldsymbol{x}_{21}]$ and $\boldsymbol{x}^2 \overset{\text{def}}{=} [\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_5, \ldots, \boldsymbol{x}_{23}]$. Eq. 3.8 denotes that the generator not only aims to fool the discriminator but also enforces the output close to the ground truth. Our final objective is

$$(G^*, D^*) = \arg\min_G \max_D \mathcal{L}_{\text{GAN}}(G, D) + \alpha \mathcal{L}_{\ell_2}(G), \tag{3.9}$$

where α is a parameter to balance these two terms. Via integrating this HCR strategy with self-attention GAN , we have the output of the reconstruction stage

$$\boldsymbol{x}' = G^*(\boldsymbol{y}, \boldsymbol{\Phi}) = \left[(\boldsymbol{x}'_1)^\mathsf{T}, \ldots, (\boldsymbol{x}'_B)^\mathsf{T}\right]^\mathsf{T}, \tag{3.10}$$

which is the desired 3D hyperspectral image.

3.1.2 Refinement Stage

The reconstruction stage can capture the spectral information of the hyperspectral image cube but it doesn't have sufficient capability to offer high quality images, especially the spatial resolution. Otherwise, an even deeper network should be used but this will require larger training datasets. To overcome this challenge, we propose the refinement stage to enhance the reconstruction quality. The input for refinement stage is a single frame instead of all spectral channels in one shot. In this manner, the network treats each spectral channel as an independent image, and it can extract the information across all spectral channels. Given the fact that the input and output images share the same structure, we

use another U-net as the basic architecture in the refinement stage; but this time we output a single frame with high quality. Since each frame is of a small size, a shallow U-net is sufficient for this task, *i.e.*, 4 times down-sampling or deconvolution in the encoder and decoder, respectively. Furthermore, we also add the *residual learning* to the input image, which has improved (1.27dB in PSNR in Table 3.2) the final results.

We pass every single frame in the hyperspectral image cube obtained by the reconstruction stage to the refinement stage. The ℓ_2 loss between the ground truth and the output of the refinement stage is used as the objective function

$$\mathcal{L}_{\ell_2}(\text{refine}) = \mathbb{E}_{\boldsymbol{x}_i, \boldsymbol{x}'_i}[\|\boldsymbol{x}_i - \boldsymbol{x}'_i\|_2], \quad \forall i = 1, \ldots, B. \tag{3.11}$$

We train the network in the reconstruction stage first and fix the parameters; then we sent the results to the refinement stage to train the second U-net. This separate training strategy is mainly due to the size difference of the data. As mentioned above, the reconstruction stage outputs the 3D hyperspectral image cube but the refinement stage processes each spectral frame independently. It is possible to train both networks jointly. However, since each batch in the reconstruction stage contains all channels of the same scene, while in the refinement stage, we hope each batch consisting of different scenes (probably at different spectral channels, too), we may need a huge memory to save these data and parameters. Limited by the GPU memory, we perform our experiments via separate training.

3.1.3 Experiments

We compare network with several state-of-the-art methods including TwIST [95], GAP-TV [96], and DeSCI [42]. We have also tried the sparse coding algorithms in [97, 98]; they perform worse than DeSCI and take even longer time to run. Similar cases exist in other algorithms [99, 100] and thus ignored here due to space limit. Both peak-signal-to-noise-ratio (PSNR) and structural similarity (SSIM) [101] are used as metrics to eval-

uate the performance. As mentioned earlier, the most recently proposed DeSCI algorithm delivers state-of-the-art results [42]. The network consistently produces high performance results and surpasses DeSCI in the “Real-Mask-in-the-Loop" (MIL) simulation data (Figs 3.7-3.6 and Table 3.1). Hereby the MIL-simulation denotes that we generate the measurement using *real masks* captured by the CASSI camera, rather than randomly generated ones. It is well known that the real captured data have noise inside and thus the problem is more challenging. On real data (we can only have a single real data with ground truth from the authors of CASSI), our network has also achieved better results than DeSCI (Figs 3.12-3.9).

Though our network is the first network developed for CASSI reconstruction for *real* data, we do compare with some other networks even they are developed for other tasks. With some modifications, we have compared network with the networks developed in [102, 103, 104] for CASSI reconstruction.

3.1.3.1 Training

All experiments are performed on a NVIDIA GTX 1080 Ti GPU. For a testing scene with size $256 \times 256 \times 24$, our framework can finish the reconstruction stage in 23ms (0.6s on CPU). In the refinement stage, every frame of the scene can be processed in parallel and finished within 10ms (0.4s on CPU). Without using the GPU, network can finish both stages on an i7 CPU within 1 second.

Figure 3.5: 16 testing scenes used in the experiments.

3.1.3.2 Data Augmentation

The data to train and validate the model is downloaded from [34]. We manually chose 80 hyperspectral images as our training data to avoid the test scenes (Fig. 3.5) and training data having the same content. Besides randomly flipping the image, we also randomly rotate, scale, and translate the training images. The original dataset have a uniform resolution of $1392 \times 1300 \times 31$ in wavelength range from 400nm to 700nm with a 10nm interval, while the real data captured by the SCI camera has 24 channels from 400nm to 700nm, but with different intervals, *i.e.*, with wavelengths: {398.62, 404.40, 410.57, 417.16, 424.19, 431.69, 439.70, 448.25, 457.38, 467.13, 477.54, 488.66, 500.54, 513.24, 526.8., 541.29, 556.78, 573.33, 591.02, 609.93, 630.13, 651.74, 674.83, 699.51}nm. To mitigate this issue, we use the *spectral interpolation* to unify the datasets to the same wavelength set as in [29]. Specifically, we perform data interpolation for every spatial location. The hyperspectral images generated by our data augmentation are of size $1392 \times 1300 \times 24$.

3.1.3.3 Training Details

We randomly crop $256 \times 256 \times 24$ patches from the data obtained by the data augmentation. The batch size is set to 20. We alternately update the parameters in G and D in the reconstruction stage; α in Eq. (3.9) is set to 200. The input of the generator is the concatenation of measurement and masks (Fig. 3.1 bottom-right). We have performed the experiments to show that this performs better (2.09dB PSNR improvement) than only input the measurement to the network in Table 3.2. During testing, we input every single channel of the hyperspectral image cube obtained by the reconstruction stage to the refinement stage. Then we collect these $B = 24$ channel high quality images as the final output result. The codes are available at *https://github.com/xinxinmiao/lambda-net*.

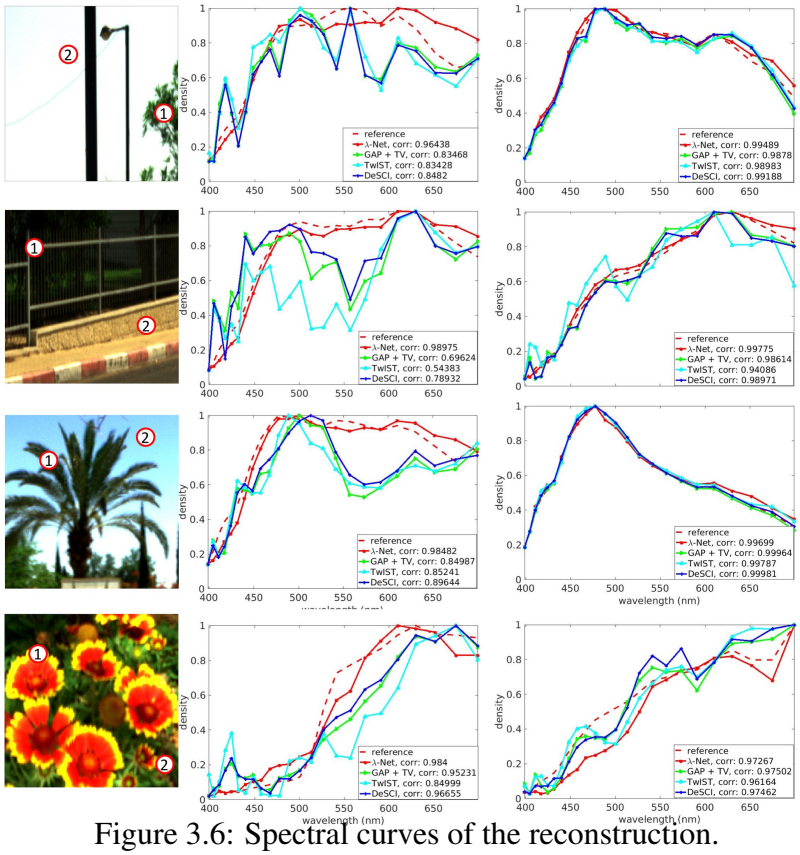


Figure 3.6: Spectral curves of the reconstruction.

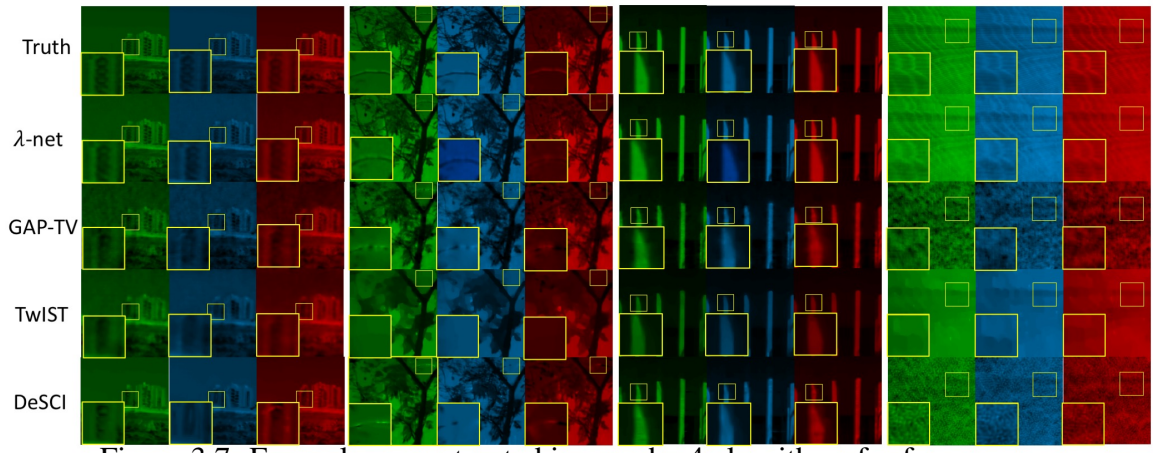


Figure 3.7: Example reconstructed images by 4 algorithms for four scenes.

3.1.3.4 “Real-Mask-in-the-Loop" Simulation Results

As mentioned above, in the MIL-simulation, we generate the measurements using the real captured mask and the hyperspectral images consist of 24 spectral frames with each of size 256×256 pixels. We have 16 testing scenes (Fig. 3.5) from the dataset [34].

Table 3.1: PSNR in dB (left entry in each cell) and SSIM (right entry) of 16 different scenes reconstructed by different algorithms.

Algorithm	network	GAP-TV	TwIST	DeSCI
Scene 1	**36.29**, **0.925**	29.48, 0.800	26.77, 0.772	31.51, 0.896
Scene 2	**30.07**, **0.929**	16.58, 0.805	13.14, 0.753	22.39, 0.806
Scene 3	**34.19**, **0.940**	21.48, 0.769	23.66, 0.738	24.92, 0.822
Scene 4	28.90, 0.899	26.49, 0.822	26.08, 0.861	**29.78**, **0.907**
Scene 5	**34.58**, **0.890**	26.63, 0.688	22.45, 0.695	29.02, 0.844
Scene 6	**28.09**, **0.858**	22.81, 0.614	20.11, 0.662	24.75, 0.797
Scene 7	**36.15**, **0.942**	24.95, 0.699	26.20, 0.753	29.68, 0.881
Scene 8	**32.64**, **0.909**	21.26, 0.695	18.38, 0.643	25.58, 0.823
Scene 9	**33.83**, **0.912**	29.94, 0.812	28.09, 0.807	32.86, 0.937
Scene 10	**28.63**, **0.877**	23.04, 0.706	20.84, 0.620	24.00, 0.748
Scene 11	**35.21**, **0.946**	24.07, 0.754	21.75, 0.785	28.19, 0.912
Scene 12	**34.77**, **0.823**	28.99, 0.758	26.75, 0.699	31.80, 0.863
Scene 13	**32.07**, **0.844**	27.57, 0.650	24.54, 0.718	30.91, 0.823
Scene 14	**33.73**, **0.869**	28.54, 0.764	26.27, 0.765	29.69, 0.852
Scene 15	**29.88**, **0.913**	25.80, 0.801	23.84, 0.765	27.45, 0.864
Scene 16	**30.54**, **0.855**	11.99, 0.293	20.50, 0.511	19.42, 0.305
average	**32.29**, **0.896**	24.35, 0.715	23.09, 0.722	27.62, 0.818

The generated measurements and masks are used to reconstruct the hyperspectral images by different algorithms. Table 3.1 lists the average PSNR and SSIM of these 16 scenes by using all four algorithms. It can be seen that in average, our network surpasses the best previous method DeSCI 4.67 dB. The only exception is Scene 4, which is a simple scene with a large area being the same white screen. This fits the rank minimization model in DeSCI and thus DeSCI offers 0.88dB higher PSNR. network performs better than DeSCI on all other scenes. Exemplar reconstructed frames of various algorithms compared with the truth are shown in Fig. 3.7. Obviously, network can provide both large-scale structures and fine details of the scene. GAP-TV usually leads to blob artifacts and TwIST provides blocky artifacts. DeSCI offers better results than GAP-TV and TwIST; however, as observed in [42], it usually leads to over-smooth reconstruction. One important metric to

evaluate the SCI algorithm is how good the spectral information they can reconstruct as different objects have different spectral information, *e.g.* sky, tree, wall, etc. We plot the spectral curves of a small region and calculate the correlation between the reconstruction and ground truth in Fig. 3.6. Compared with other methods, network provides higher correlation values for different objects. This clearly demonstrates that network can extract more spectral information than other methods.

To quantitatively investigate different blocks of our proposed network, we performed experiments with partial components in network, *e.g.*, without GAN, without self-attention, with results summarized in Table 3.2. It can be seen that all components play important roles in our network; *e.g.*, without GAN, the results degraded 2.81dB in PSNR; without self-attention, the results degraded 3.52dB in PSNR, and without the refinement stage, the results degraded 1.62dB in PSNR. As mentioned before, masks contain useful information, and thus using masks along with the measurement improved the results by 2.09dB in PSNR. Furthermore, residual learning in the refinement U-net has led to 1.27dB improvement in PSNR and HCR imporves the result for 0.48dB.

Table 3.2: Comparison using different components of the model.

Reconstruction stage	U-net [39]	×	×	×	×	×	×	√	×
	Deep U-net1	√	√	√	√	√	√	×	√
	GAN	√	√	×	√	√	√	√	√
	Self-attention	√	√	√	×	√	√	√	√
	HCR	√	×	√	√	√	√	√	√
Refinement stage	U-net2	√	√	√	√	×	√	√	√
	residual learning	√	√	√	√	×	√	√	×
inputs	measurement+masks	√	√	√	√	√	×	√	√
	measurement	×	×	×	×	×	√	×	×
result PNSR result SSIM		32.29 0.896	31.81 0.882	29.48 0.860	28.77 0.854	30.67 0.873	30.20 0.866	30.37 0.870	31.02 0.878

As mentioned before, we have also compared our network with other networks, with our modifications for CASSI reconstruction. The results are summarized in Table 3.3, where we can observe that network provides significant better results than other networks.

Table 3.3: Compare with other deep networks.

Network	Simu PSNR	Simu SSIM	Real PSNR
network	32.29	0.896	25.59
[4]	27.42	0.750	21.42
[5]	26.78	0.735	21.09
[6]	29.07	0.836	23.77

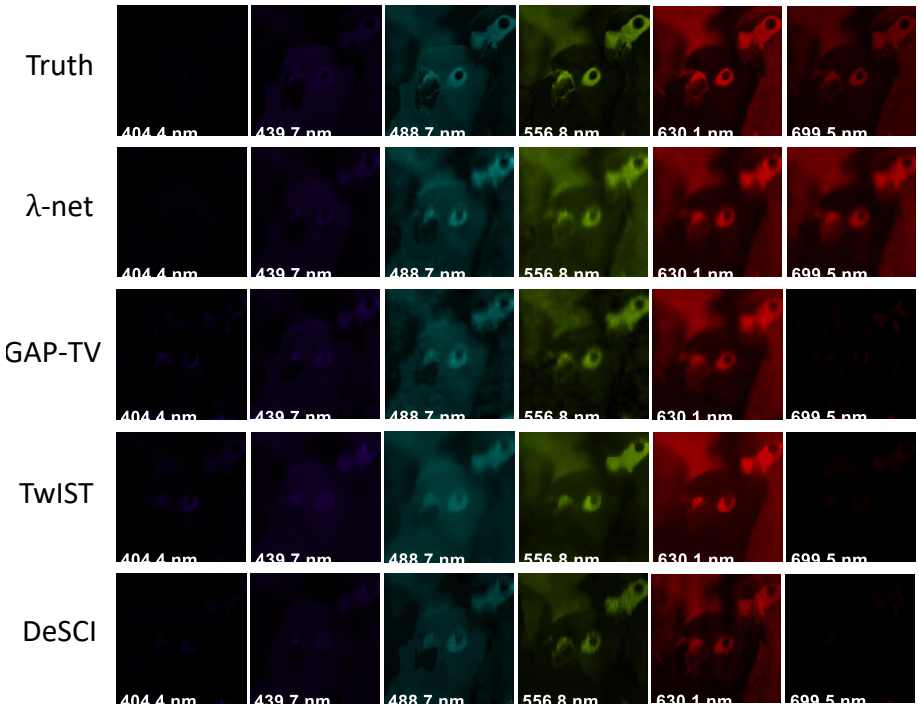


Figure 3.8: Real data results: reconstructed bird data from measurement captured by the real camera.

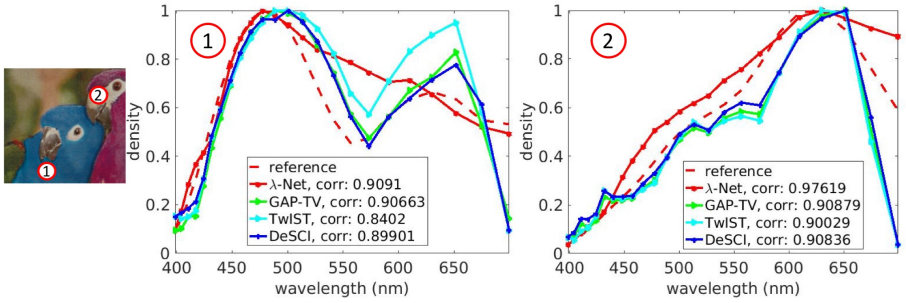


Figure 3.9: Real data results: reconstructed spectral of the bird data from measurement captured by the real SCI camera.

3.1.3.5 Real Data Results

The bird measurement data is captured by the CASSI system [29]. consisting of 24 spectral frames with each of size 1021×703 pixels. Due to the limitation of GPU memory, we used 416×416 pixels to perform our experiments[1]. In Fig. 3.12, we visualize the reconstruction results of 6 channels using 4 algorithms. We can see that network can provide marginally better (0.4dB) results than DeSCI and about 1dB higher PSNR than GAP-TV and TwIST. Notably, only network can reconstruct the last frame at wavelength 699.5nm. Exemplar spectral curves are shown in Fig. 3.9. Owing to the mismatch between the training dataset and this real data, the spectra are not perfect; even this, network can still offer higher or comparable correlation values with other three algorithms.

As mentioned before, we only have one real data, *i.e.*, the bird data, with ground truth captured by CASSI. To further verify the universality of our network, we have modified the network to the video CS system [105]. The results are comparable with DeSCI.

3.2 Reconstruct Video Images from a Snapshot Measurement

Another representative application of SCI system is video compressive imaging. We can simply update the current model to reconstruct video frames from a snapshot measurement. The first update we have done is removing the self-attention module. The reason is for every single channel in the same feature maps they share the same attention map. It is helpful because the object in the hyperspectral images are not moving. However, the object is moving in videos throught the time axis, so every single channel should not have the

[1]It is possible to train multiple networks for different regions of the large area, since the mask values for different places are different. However, the training takes too long and multiple GPUs are required, which is beyond our capability. We believe this 416×416 region can demonstrate the performance of our proposed network.

same attention. That's why we remove the self-attention module for video images reconstruction.

3.2.1 Consistent loss

Since we have the mask that have used in our camear, it is reasonable to use this information as much as possible in our framework. After we generate the video we want, we can then generate the measurement as the same process happened in the camear. The generated measurement should match the measurement that captured by our camera. So we also add one consistent loss to make the generated measurement same as the captured measurement.

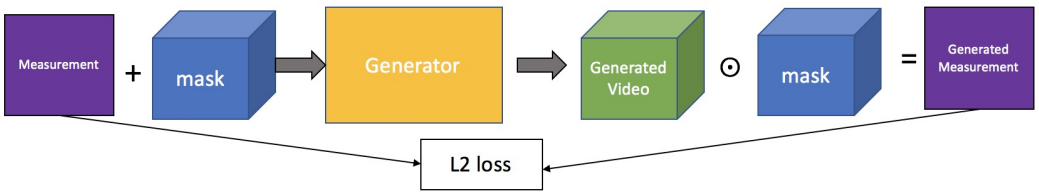


Figure 3.10: Structure for consistent loss.

3.2.2 Total variation loss

In image processing, total variation denoising, also known as total variation regularization, is often used for removing noise from the image. Our refinement network is designed to denoising for every single channel image in the framework which is also proved can improve the PSNR for the reconstructed hyperspectral images. So, besides the L2 loss the total variation loss will be also added to the generator to reduce the noise for the generated video.

3.2.3 Experiment

The same as reconstruct hyperspectral images, we compare network with several state-of-the-art methods including TwIST [95], GAP-TV [96], and DeSCI [42]. And also we have two parts for the experiment simulation and real data. For simulation Peak-signal-to-noise-ratio (PSNR) is used as metrics to evaluate the performance. On real data (we can only have a single real data), since we don't have the groundtruth for the framese. We only visulize the results.

The training data we used is called High-Speed-Video Slow-Motion Demonstrations which can be downloaded from https://high-speed-video.colostate.edu/. There are 200 videos are used to train our deep network. For the testing data, two datasets, namely Kobe and Traffic, used in [42] are also used in our simulation. On the other hand, we also pick up one testing video from the High-Speed-Video Slow-Motion Demonstrations. The same as [42], for the traffic dataset we employ related data to train the network.

3.2.3.1 Simulation Results

As mentioned above, we generate the measurements using shifted mask and the video clip consist of 8 frames with each of size 256×256 pixels. We have 3 testing scenes. The generated measurements and masks are used to reconstruct the frames by different algorithms. Table 3.4 lists the PSNR of these 3 scenes by using all four algorithms. It shows that out model gives best result for 2 testing scenes. For that 'Kobe' dataset, our model is worse than DeSCI, the main reason is our training dataset is totally different with the basketball testing scene. Exemplar reconstructed frames are shown in Fig. 3.11. From left to right are truth, our result, DeSCI, GAP-TV. Our result is the best one.

Table 3.4: PSNR in dB (left entry in each cell) 3 different scenes reconstructed by different algorithms.

Algorithm	network	Desci	GAP-TV	MMLE-MFA	MMLE-GMM	GMM-TP
Traffic	29.04	28.72	20.89	22.66	25.68	25.08
Kobe	25.75	33.25	26.45	24.63	27.33	24.47
Ballon	36.10	34.94	27.77	28.51	31.72	29.31

Figure 3.11: Reconstruction result for one of the simulation data

3.2.3.2 Real data Results

Because our camera caputre the measurement directly for the scene, we don't have the ground truth for the real data. We only visuliza the reconstruction result for the result in Fig 3.12.

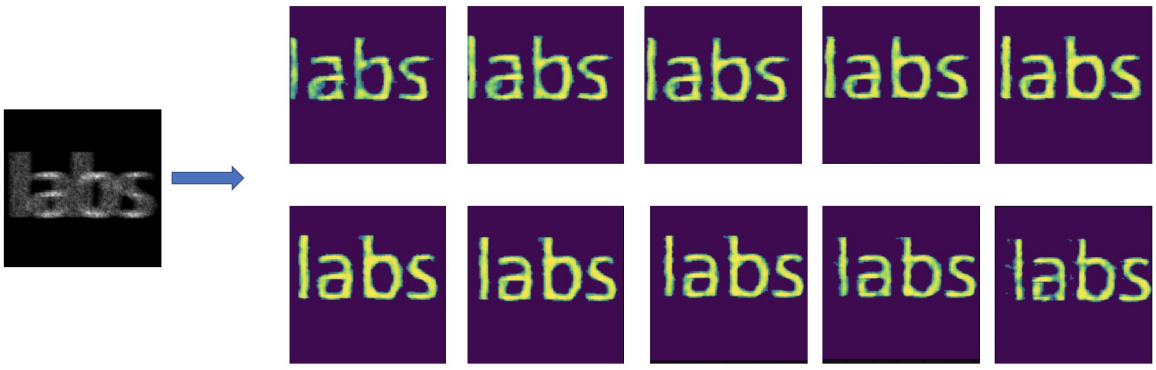


Figure 3.12: Result for real data

CHAPTER 4

Conclusion

Many challenges for high dimensional images processing remain unsolved, even though some existing works have achieved good performance in different applications and it has attracted large attentions recently. The deep learning methods became very popular recently because it's feature representation ability. This paper focus on how to design the efficient deep learning networks for two computer vision applications, facial landmarks detection and imges reconstruction for snapshot compressive imaging.

In this paper, we firstly propose the direct shape regression network (DSRN) for end-to-end facial landmarks detection in a unified framework. Specifically, by deploying doubly convolutional layer and by using the Fourier feature pooling layer proposed in this paper, DSRN efficiently constructs strong representations to disentangle highly nonlinear relationships between images and shapes; by incorporating a linear layer of low-rank learning, DSRN effectively encodes correlations of landmarks to improve performance. DSRN leverages the strengths of kernels for nonlinear feature extraction and neural networks for structured prediction, and provides the first end-to-end learning architecture for direct face alignment. All empirical results demonstrate that DSRN consistently produces high performance and in most cases surpasses state-of-the-art.

In addition, we also address the challenging problem in snapshot compressive imaging: the slow reconstruction. Inspired by the recent advances of deep learning, especially the emerging generative models, we have built a two-stage reconstruction network to recover the images from a snapshot measurement. By integrating U-net into the GAN framework, we have incorporated the nonlocal similarity in the images into the reconstruction

network, thus have improved the performance of our model. The hierarchical channel reconstruction has been proposed to decompose the hard problem into several easier tasks. The experiment results proved that HCR can further improve the performance. To further enhance the quality of reconstructed images, we have adapted another small U-net with residual learning to refine the results of the first stage. By processing each frame independently, the parameters in this second U-net have decreased dramatically and thus it is easy to train. The quality of reconstructed images has improved significantly due to this refinement stage. Our proposed network has been verified by the real data captured by the compressive camera. It not only achieves better results than the current state-of-the-art, but also finishes the reconstruction in a short time. It is expected to use the compressive camera with our network to build an end-to-end video-rate 3D imaging system, while enjoying the benefits of low cost and low bandwidth.

REFERENCES

[1] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 363–378.

[2] Z. Lan, Y. Zhu, A. G. Hauptmann, and S. Newsam, "Deep local video feature for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1–7.

[3] Y. Zhu and S. Newsam, "Densenet for dense flow," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 790–794.

[4] ——, "Efficient action detection in untrimmed videos via multi-task learning," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 197–206.

[5] Y. Wang and M. Hoai, "Improving human action recognition by non-action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2698–2707.

[6] V. Tran, Y. Wang, and M. Hoai, "Back to the future: Knowledge distillation for human action anticipation," *arXiv preprint arXiv:1904.04868*, 2019.

[7] Y. Wang, V. Tran, G. Bertasius, L. Torresani, and M. Hoai, "Attentive action and context factorization," *arXiv preprint arXiv:1904.05410*, 2019.

[8] Y. Wang and M. Hoai, "Pulling actions out of context: Explicit separation for effective combination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7044–7053.

[9] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8856–8865.

[10] J. Yang, W. An, S. Wang, X. Zhu, C. Yan, and J. Huang, "Label-driven reconstruction for domain adaptation in semantic segmentation," *arXiv preprint arXiv:2003.04614*, 2020.

[11] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018, pp. 103–110.

[12] S. Wang, Z. Xu, C. Yan, and J. Huang, "Graph convolutional nets for tool presence detection in surgical videos," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 467–478.

[13] J. Yao, S. Wang, X. Zhu, and J. Huang, "Imaging biomarker discovery for lung cancer survival prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 649–657.

[14] S. Wang, A. Raju, and J. Huang, "Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 620–623.

[15] Z. Xu, S. Wang, Y. Li, F. Zhu, and J. Huang, "Prim: An efficient preconditioning iterative reweighted least squares method for parallel brain mri reconstruction," *Neuroinformatics*, vol. 16, no. 3-4, pp. 425–430, 2018.

[16] F. Zheng, X. Miao, and H. Huang, "Fast vehicle identification via ranked semantic sampling based embedding," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence. AAAI Press*, 2018, pp. 3697–3703.

[17] Q. Jin, L. Yang, and Z. Liao, "Adabits: Neural network quantization with adaptive bit-widths," *arXiv preprint arXiv:1912.09666*, 2019.

[18] ——, "Towards efficient training for neural network quantization," *arXiv preprint arXiv:1912.10207*, 2019.

[19] X. Miao, X. Zhen, X. Liu, C. Deng, V. Athitsos, and H. Huang, "Direct shape regression networks for end-to-end face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5040–5049.

[20] L. Yue, X. Miao, P. Wang, B. Zhang, X. Zhen, and X. Cao, "Attentional alignment networks." in *BMVC*, vol. 2, no. 6, 2018, p. 7.

[21] G. Toderici, G. Passalis, S. Zafeiriou, G. Tzimiropoulos, M. Petrou, T. Theoharis, and I. A. Kakadiaris, "Bidirectional relighting for 3d-aided 2d face recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2721–2728.

[22] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 3, pp. 526–534, 2012.

[23] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[26] S. Zhai, Y. Cheng, Z. M. Zhang, and W. Lu, "Doubly convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 1082–1090.

[27] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[28] M. E. Gehm, R. John, D. J. Brady, R. M. Willett, and T. J. Schulz, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Optics Express*, vol. 15, no. 21, pp. 14 013–14 027, 2007.

[29] A. Wagadarikar, R. John, R. Willett, and D. J. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Applied Optics*, vol. 47, no. 10, pp. B44–B51, 2008.

[30] A. Wagadarikar, N. Pitsianis, X. Sun, and D. Brady, "Video rate spectral imaging using a coded aperture snapshot spectral imager," *Optics Express*, vol. 17, no. 8, pp. 6368–6388, 2009.

[31] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, Sept 2017.

[32] X. Yuan and Y. Pu, "Parallel lensless compressive imaging via deep convolutional neural networks," *Optics Express*, vol. 26, no. 2, pp. 1962–1977, Jan 2018.

[33] "Hyperspectral and color imaging," https://sites.google.com/site/hyperspectralcolorimaging/dataset/general-scenes, accessed: 2018-11-05.

[34] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural rgb images," in *European Conference on Computer Vision*. Springer, 2016, pp. 19–34.

[35] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "λ-net: Reconstruct hyperspectral images from a snapshot measurement," in *IEEE/CVF Conference on Computer Vision (ICCV)*, vol. 1, 2019.

[36] ——, "l-net: Reconstruct hyperspectral images from a snapshot measurement," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4059–4069.

[37] X. Miao, X. Yuan, and P. Wilford, "Deep learning for compressive spectral imaging," in *Digital Holography and Three-Dimensional Imaging*. Optical Society of America, 2019, pp. M3B–3.

[38] X. Miao, X. Yuan, Y. Pu, and V. Athitsos, "Supplementary material for ?λ-net: Reconstruct hyperspectral images from a snapshot measurement?"

[39] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241, (available on arXiv:1505.04597 [cs.CV]). [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a

[40] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[42] Y. Liu, X. Yuan, J. Suo, D. Brady, and Q. Dai, "Rank minimization for snapshot compressive imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. in press, 2018.

[43] J. Zhang, M. Kan, S. Shan, and X. Chen, "Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders," in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3428–3437.

[44] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[45] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[46] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[47] S. Zhu, C. Li, C.-C. Loy, and X. Tang, "Unconstrained face alignment via cascaded compositional learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3409–3417.

[48] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

[49] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4998–5006.

[50] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2729–2736.

[51] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3476–3483.

[52] H. Liu, J. Lu, J. Feng, and J. Zhou, "Two-stream transformer networks for video-based face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[53] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[54] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage reinitialization for high performance facial landmark detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[55] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[56] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4177–4187.

[57] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," *arXiv preprint arXiv:1703.07332*, 2017.

[58] ——, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," *arXiv preprint arXiv:1703.00862*, 2017.

[59] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," *arXiv preprint arXiv:1611.05396*, 2016.

[60] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.

[61] B. Gu, M. Xin, Z. Huo, and H. Huang, "Asynchronous doubly stochastic sparse kernel learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[62] S. Bochner, *Lectures on Fourier Integrals.(AM-42)*. Princeton University Press, 2016, vol. 42.

[63] A. Sinha and J. C. Duchi, "Learning kernels with random features," in *Advances in Neural Information Processing Systems*, 2016, pp. 1298–1306.

[64] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *ICASSP*, 2013.

[65] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2144–2151.

[66] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[67] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[68] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*. Springer, 2012, pp. 679–692.

[69] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*, vol. 964, 1999, pp. 965–966.

[70] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 50–58.

[71] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[72] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 918–930, 2016.

[73] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1944–1951.

[74] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[75] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[76] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3659–3667.

[77] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[78] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[79] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1851–1858.

[80] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *European Conference on Computer Vision*. Springer, 2014, pp. 1–16.

[81] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment via regressing local binary features," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1233–1245, 2016.

[82] X. Yu, F. Zhou, and M. Chandraker, "Deep deformation network for object landmark localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 52–70.

[83] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[84] P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," *Optics express*, vol. 21, no. 9, pp. 10 526–10 545, 2013.

[85] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, February 2006.

[86] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[87] S. Jalali and X. Yuan, "Compressive imaging via one-shot measurements," in *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[88] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, 2014, pp. 2672–2680.

[89] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[90] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[91] K. Nazeri, E. Ng, and M. Ebrahimi, "Image colorization using generative adversarial networks," in *International Conference on Articulated Motion and Deformable Objects*. Springer, 2018, pp. 85–94.

[92] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[93] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, cite arxiv:1411.1784. [Online]. Available: http://arxiv.org/abs/1411.1784

[94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[95] J. Bioucas-Dias and M. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Processing*, vol. 16, no. 12, pp. 2992–3004, December 2007.

[96] X. Yuan, "Generalized alternating projection based total variation minimization for compressive sensing," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 2539–2543.

[97] L. Wang, Z. Xiong, G. Shi, F. Wu, and W. Zeng, "Adaptive nonlocal sparse representation for dual-camera compressive hyperspectral imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 10, pp. 2104–2111, Oct 2017.

[98] L. Wang, Z. Xiong, H. Huang, G. Shi, F. Wu, and W. Zeng, "High-speed hyperspectral video acquisition by combining nyquist and compressive sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[99] J. Yang, X. Liao, X. Yuan, P. Llull, D. J. Brady, G. Sapiro, and L. Carin, "Compressive sensing by learning a Gaussian mixture model from measurements," *IEEE Transaction on Image Processing*, vol. 24, no. 1, pp. 106–119, January 2015.

[100] X. Yuan, T.-H. Tsai, R. Zhu, P. Llull, D. J. Brady, and L. Carin, "Compressive hyperspectral imaging with side information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 964–976, September 2015.

[101] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[102] S. Koundinya, H. Sharma, M. Sharma, A. Upadhyay, R. Manekar, R. Mukhopadhyay, A. Karmakar, and S. Chaudhury, "2d-3d cnn based architectures for spectral reconstruction from rgb images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[103] H. Li, Z. Xiong, Z. Shi, L. Wang, D. Liu, and F. Wu, "Hsvcnn: Cnn-based hyperspectral reconstruction from rgb videos," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 3323–3327.

[104] Z. Shi, C. Chen, Z. Xiong, D. Liu, and F. Wu, "Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[105] P. Llull, X. Liao, X. Yuan, J. Yang, D. Kittle, L. Carin, G. Sapiro, and D. J. Brady, "Coded aperture compressive temporal imaging," *Optics Express*, vol. 21, no. 9, pp. 10 526–10 545, 2013.