9-2017

Exploring the Internal Statistics: Single Image Super-Resolution, Completion and Captioning

Yang Xian

The Graduate Center, City University of New York

How does access to this work benefit you? Let us know!

Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Computer Sciences Commons

Recommended Citation

https://academicworks.cuny.edu/gc_etds/2277

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
Exploring the Internal Statistics: Single Image
Super-Resolution, Completion and Captioning

by

Yang Xian

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2017
This manuscript has been read and accepted by the Graduate Faculty in Computer Science in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Professor YingLi Tian

__________________________________________
Date                                      Chair of Examining Committee

Professor Robert M. Haralick

__________________________________________
Date                                      Executive Officer

Professor YingLi Tian
Professor Ioannis Stamos
Professor Zhigang Zhu
Dr. Lisa M. Brown
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

EXPLORING THE INTERNAL STATISTICS: SINGLE IMAGE SUPER-RESOLUTION, COMPLETION AND CAPTIONING

by

YANG XIAN

Adviser: Professor YingLi Tian

Image enhancement has drawn increasingly attention in improving image quality or interpretability. It aims to modify images to achieve a better perception for human visual system or a more suitable representation for further analysis in a variety of applications such as medical imaging, remote sensing, and video surveillance. Based on different attributes of the given input images, enhancement tasks vary, e.g., noise removal, deblurring, resolution enhancement, prediction of missing pixels, etc. The latter two are usually referred to as image super-resolution and image inpainting (or completion).

Image super-resolution and completion are numerically ill-posed problems. Multi-frame-based approaches make use of the presence of aliasing in multiple frames of the same scene. For cases where only one input image is available, it is extremely challenging to estimate the unknown pixel values. In this dissertation, we target at single image super-resolution and completion by exploring the internal statistics within the input image and across scales. An internal gradient similarity-based single image super-resolution algorithm is first presented. Then we demonstrate that the proposed framework could be naturally extended to accomplish super-resolution and completion simultaneously. Afterwards, a hybrid learning-based single image super-resolution approach is proposed to benefit from both external and internal statistics. This framework hinges on image-level hallucination from externally learnt regression models as well as gradient level pyramid self-awareness for edges and textures.
refinement. The framework is then employed to break the resolution limitation of the passive microwave imagery and to boost the tracking accuracy of the sea ice movements. To extend our research to the quality enhancement of the depth maps, a novel system is presented to handle circumstances where only one pair of registered low-resolution intensity and depth images are available. High quality RGB and depth images are generated after the system. Extensive experimental results have demonstrated the effectiveness of all the proposed frameworks both quantitatively and qualitatively.

Different from image super-resolution and completion which belong to low-level vision research, image captioning is a high-level vision task related to the semantic understanding of an input image. It is a natural task for human beings. However, image captioning remains challenging from a computer vision point of view especially due to the fact that the task itself is ambiguous. In principle, descriptions of an image can talk about any visual aspects in it varying from object attributes to scene features, or even refer to objects that are not depicted and the hidden interaction or connection that requires common sense knowledge to analyze. Therefore, learning-based image captioning is in general a data-driven task, which relies on the training dataset. Descriptions in the majority of the existing image-sentence datasets are generated by humans under specific instructions. Real-world sentence data is rarely directly utilized for training since it is sometimes noisy and unbalanced, which makes it ‘imperfect’ for the training of the image captioning task. In this dissertation, we present a novel image captioning framework to deal with the uncontrolled image-sentence dataset where descriptions could be strongly or weakly correlated to the image content and in arbitrary lengths. A self-guiding learning process is proposed to fully reveal the internal statistics of the training dataset and to look into the learning process in a global way and generate descriptions that are syntactically correct and semantically sound.
Acknowledgments

This dissertation would not have been possible without the support and guidance from many people during the five wonderful years of my PhD study.

First and foremost, I would like to express my deepest gratitude to Prof. YingLi Tian, for being such a wonderful mentor and friend. I am thankful to her for the tremendous guidance, encouragement, and support, for the patience that she had during some of the difficult times in my research, for the amount of independence she gave me to try new ideas, and for the confidence she has showed in me. I am sure her insight, intuition, and elegance in problem solving will inspire me in my future endeavors.

I am also honored to have Prof. Ioannis Stamos, Prof. Zhigang Zhu, and Dr. Lisa M. Brown on my dissertation committee. Their insightful comments and suggestions are very much appreciated. My gratitude goes to Dr. Yunqing Wang, my intern mentor in Google during the summer of 2016.

I would like to thank all my colleagues over the years: Dr. Shizhi Chen, Dr. Chucai Yi, Dr. Xiaodong Yang, Dr. Chenyang Zhang, Xuejian Rong, Yuancheng Ye, Dr. Zisis I. Petrou, Jingya Liu, Longlong Jing, and Wendy Fernandez.

I dedicate this dissertation to my beloved family members. My infinite gratitude goes to my parents for their unconditioned love and support, and to my husband and best friend, Yuan, whose love and friendship make everything meaningful.
Contents

1 Introduction ............................................. 1
  1.1 Image Super-Resolution .................................. 2
    1.1.1 Interpolation-based Super-Resolution .................. 4
    1.1.2 Reconstruction-based Super-Resolution .................. 6
    1.1.3 Learning-based Super-Resolution ....................... 7
  1.2 Image Completion ...................................... 15
  1.3 Image Captioning ....................................... 17
    1.3.1 Methods ........................................... 18
    1.3.2 Datasets ........................................... 20

2 Internal Example-based Image Enhancement ........... 22
  2.1 Internal Gradient Similarity-based Super-Resolution .... 22
    2.1.1 Framework Details ................................... 22
    2.1.2 Why Gradient Level Works Better? ..................... 26
    2.1.3 Internal vs. External Statistics ....................... 30
    2.1.4 Experimental Results ................................ 37
  2.2 Internal Example-based Image Enhancement ............ 46
    2.2.1 Framework Details ................................... 47
    2.2.2 Experimental Results ................................ 50
CONTENTS

2.3 Discussions ............................................................... 54

3 Hybrid Example-based Super-Resolution and its Application 56
  3.1 Hybrid Example-based Super-Resolution ............................ 57
    3.1.1 Framework Details .................................................. 58
    3.1.2 Experimental Results .............................................. 66
  3.2 Super-Resolved Fine Scale Sea Ice Motion Tracking ............... 70
    3.2.1 Data ................................................................. 75
    3.2.2 Framework Details ................................................ 75
    3.2.3 Experimental Results ............................................ 80
  3.3 Discussions ............................................................ 92

4 Resolution Enhancement in Single Depth Map and Aligned Image 93
  4.1 Framework Details ...................................................... 97
    4.1.1 Depth Completion ................................................ 97
    4.1.2 Super Resolution .................................................. 100
  4.2 Experimental Results .................................................. 101
    4.2.1 Implementation Details ......................................... 102
    4.2.2 Quantitative Comparison ......................................... 103
    4.2.3 Qualitative Comparison .......................................... 104
  4.3 Discussions ............................................................ 107

5 Self-Guiding Multimodal LSTM - when we do not have a perfect training dataset for image captioning 109
  5.1 Framework Details ....................................................... 112
    5.1.1 Captioning with m-LSTM ........................................ 115
    5.1.2 Captioning with sg-LSTM ....................................... 117
## CONTENTS

5.2 Experimental Results .................................................. 121
   5.2.1 FlickrNYC Dataset ................................................. 121
   5.2.2 Implementation Details ......................................... 122
   5.2.3 Experimental Evaluations ..................................... 123
5.3 Discussions .............................................................. 126

6 Conclusions ............................................................... 127

Appendices ................................................................. 128

A Publications ............................................................. 129

Bibliography ................................................................. 131
List of Tables

2.1 Results of the percentage of images which have larger SSIM when general external statistics is introduced at three different input sizes. .......................... 33

2.2 Results of the percentage of images which have larger SSIM when class-specific external statistics is introduced at 25 different classes in the UIUC Texture Dataset [66]. .................................................................................................................. 36

2.3 IFC comparisons with peer super-resolution methods on SET5 [9]. .............. 45

3.1 Comparison of the proposed approach with recent state-of-the-art methods in SET5 [9], SET14 [145], and BSDS [79] in terms of average PSNR (dB). Our results outperform other methods in all three datasets. ................................. 67

3.2 Quantitative evaluation of the NN and SR drift vectors compared with the vectors from the original AMSR2 images, for all image pairs individually and aggregated. “δx” and “δy” indicate the drifts on the vertical and horizontal axes, respectively. “Samples” stands for the total number of drift vectors compared, “RSE” for the relative squared error, “RMSE” the root mean squared error in km, “MAE” the mean absolute error in km, and “P” the Pearson correlation coefficient. .......................................................... 87
3.3 Quantitative evaluation of the errors of the NN and SR drift vectors compared with the vectors from the original AMSR2 images, for all image pairs individually and aggregated. \( \mu_x \) and \( \mu_y \) stand for the mean error values in the \( x \) and \( y \) directions, respectively, whereas \( \sigma_x \) and \( \sigma_y \) for the respective standard deviation.

4.1 Comparison of the proposed approach with other depth super-resolution schemes over 14 depth maps in dataset [99] measured in terms of FSIM [146] under the magnification factor of 4. The best performance in FSIM for each image is marked in bold.

4.2 Comparison of the proposed approach with peer methods for intensity images in Middlebury Stereo datasets [20, 46, 98, 99] measured in terms of average FSIM [146] under the scaling factor of 4. The best performance in FSIM for each dataset is marked in bold.

5.1 Statistics of image distribution based on the description lengths in FlickrNYC dataset.

5.2 Numerical results of the proposed framework compared with other methods based on the testing images in \( data_t \).
List of Figures

1.1 Patch recurrence within and across scales of a single image [38]. Source patches in $I$ are found in different locations and in other scales of $I$ (solid-marked squares). The high-resolution corresponding parent patches (dashed-marked squares) provide an indication of what the unknown high-resolution correspondences of the source patches might look like. ......................... 14

2.1 Flowchart of the proposed across-scale gradient similarity-based super-resolution algorithm. After calculating the gradients of the input low-resolution image in horizontal and vertical directions (represented as $L_x$ and $L_y$), for each gradient patch, its top $k$ most similar patches are searched within the corresponding gradient patch pool. Patch pool $\varphi_x$ is composed of all gradient patches in the downsampled version of $L_x$ (represented as $LL_x$). $\varphi_y$ is built up in a similar manner utilizing $L_y$. After constructing the high-resolution gradients $H_x$ and $H_y$, the output image is restored based on them and the input image $L$. . . . 23
2.2 Demonstration of the feasibility to restore the high-resolution image from accurate horizontal and vertical gradients and the corresponding low-resolution image using the proposed reconstruction framework. (a) Experiments over the ground-truth images and the reconstructed images based on 100 natural images in database BSDS [79] at different scales and weights (λ). Left: pixel-wise averaged error; Middle: PSNR (dB); Right: SSIM. (b)-(g) illustrate the reconstruction results on images “child” and “smile” along with the ground-truth images and the corresponding difference maps. Error mainly exists along image boundaries. .......................................................... 28

2.3 Comparisons of the super-resolution results (×4) in the image level and the gradient level (with reconstruction) respectively using the same framework. Three examples are presented and the ground-truth images are included for two cases shown in the top and middle rows. The ground-truth image is not available for the example in the bottom row. In all three cases, it is clearly demonstrated that the results produced in the gradient level reveal sharper edges and more natural textures. ............................ 31

2.4 Example images of all the 25 classes in the UIUC Texture Dataset [66] (sorted in descending order measured in “percentage of images with larger SSIM with class-specific external statistics”. Refer to Table 2.2 for more details.) .... 35

2.5 Super-resolution of image “comic” (×8). (a) Input low-resolution image; (b) Generated high-resolution result. Super-resolution of image “butterfly” (×16). (c) Input low-resolution image; (d) Generated high-resolution result. Our method synthesizes fine details and restores clear edges. For a better presentation, the input image is upscaled to the target resolution utilizing nearest-neighbor interpolation. .......................................................... 40
2.6 Super-resolution of image “child” (×4). Our proposed method successfully reconstructs the knit textures in the hat region and maintains sharp and natural eyes, facial and lip contours so that the hat is actually above the face, not vice versa as observed in peer results. ........................................ 41

2.7 Super-resolution of image “chip” (×4). Our proposed method retains clean and sharp structural edges. ......................................................... 42

2.8 Super-resolution of images “baboon”, “Lenna”, and “pepper” (×3). In all three images, our proposed method generates results with clear edges and realistic textures closest to the ground-truth images. ................................. 43

2.9 Super-resolution of images “mushroom”, “flower”, “car”, “fish”, and “girl” (×4). Our proposed method generates results with both sharp edges and natural textures. ................................................................. 44

2.10 Flowchart of the proposed internal example-based image enhancement framework. Given an input low-resolution image with a mask indicating the missing region(s), the low-resolution gradients are upsampled with completion embedded. The high-resolution gradients, along with the completed low-resolution image, are fed into an energy function to reconstruct the final high-resolution image. ......................................................... 47

2.11 Enhancement results of image “fish” (×2) with two different masks. Left: input low-resolution images with missing regions (marked in white). Right: generated high-resolution results with clear contours and natural textures. 51

2.12 Enhancement result of image “snow” (×2). Left: input low quality image with missing region (marked in white). Right: output high quality images after enhancement. ......................................................... 52
2.13 Enhancement results of 4 images in BSDS dataset (×2) with different masks. Left: input low quality images with missing region(s) (marked in white). Right: output high quality images after enhancement. 53

3.1 Super-resolution result of image “face” (×2). (a) High-resolution image generated by the proposed super-resolution method. (b) Ground-truth image. The proposed hybrid example-based super-resolution approach reconstructs natural and realistic edges and textures close to the original ground-truth image. 57

3.2 Flowchart of the hybrid example-based super-resolution method. Given a low-resolution image, a proxy high-resolution image is constructed through the regression models trained via an external dataset. The input feature space is modeled with GMM to ensure a targeted learning. With the proxy image, its gradients are refined using gradients of the input image. The refined high-resolution gradients are then integrated into the reconstruction framework to recover the final output image. 58

3.3 Variance distributions of 20,000 patches with size 7 × 7 extracted from BSDS dataset. Given a patch, the larger the variance is, the less frequent it tends to appear within the dataset. 62

3.4 (a) Illustration of the gradient-level coarse-to-fine self-awareness procedure. Gradients of the proxy image $H_{\{x,y\}}^p$ are downsampled to $M_{\{x,y\}}^p$ which are refined with the gradients $L_{\{x,y\}}$ of the input image. Gradients with darker frames represent the corresponding refined results of the ones with lighter frames. Afterwards, $H_{\{x,y\}}^p$ is refined with $M_{\{x,y\}}$. Please refer to text for details. (b) Difference map between $H_{\{x\}}$ and $H_{\{x\}}^p$. (c) Difference map between $H_{\{y\}}$ and $H_{\{y\}}^p$. 63
3.5 Comparisons measured in average PSNR (dB) between the proxy images (marked in blue) and the corresponding final output (marked in red) recovered from self-refinement and reconstruction in datasets BSDS [79], SET5 [9], and SET14 [145] (scaling factor ×4). There is an obvious boost in PSNR after performing self-awareness and reconstruction for all three datasets. 65

3.6 Super-resolution of image “Lenna” (×4). Zoom-ins clearly indicates that the proposed hybrid example-based super-resolution framework reconstructs the hat contours with minimal artifacts while other methods suffer from blurring, jaggies, or aliasing artifacts. 68

3.7 Super-resolution of image “snow” (×4). It is a challenging task to reconstruct the rattan textures. Results generated by bicubic interpolation and [142] oversmooth the textures. Deformed patterns exist in [103] (squared textures) and [139] (discontinuities). [120] fails to recover several edges as shown in the circled zoom-in and [24] oversharpen the edges. Our result best reconstructs the details. 69

3.8 Daily composite AMSR2 image at 36.5 GHz from January 1, 2013, projected on a polar stereographic grid. The latitude-longitude coordinates of the image corners are: upper-left (30.98°, 168.35°), upper-right (31.37°, 102.34°), lower-left (33.92°, 279.26°), lower-right (34.35°, 350.03°) [87]. 73

3.9 Super-resolution of the passive microwave imagery (×4). From left to right, the columns represent zoom-ins of the ground-truth image, the low-resolution image, and the generated high-resolution image. It is clearly indicated that the hybrid example-based super-resolution approach reconstructs clearer ice contours consistent with the ground-truth instances. 81
3.10 Estimated motion vectors from the first pair (January 1-2, 2013) of images, for each of the original AMSR2 (Orig.), low-resolution nearest-neighbor (NN), and super-resolved (SR) image sets. ........................................ 82

3.11 Close-up look for the main Arctic region around the North Pole of the estimated motion vectors from the first pair (January 1-2, 2013) of images, for the (a) original AMSR2 (Orig.), (b) low-resolution nearest-neighbor interpolated (NN), and (c) super-resolved (SR) image sets. .............................. 84

3.12 Scatterplots of NN and SR drift estimates (in km) in the two axes compared with the vectors from the original images considered as the ground-truth, for the image pairs on January 1 and 2. (a) NN drift in the \( x \)-axis (vertical); (b) NN drift in the \( y \)-axis (horizontal); (c) SR drift in the \( x \)-axis; (d) SR drift in the \( y \)-axis. In addition to the drift data, the 1-by-1 ideal match line is drawn as a dashed line, as well as the least squares linear regression fit line as a solid one. ................................................................. 85

3.13 Error distributions from the (a) NN and (b) SR drift estimates in the two axes compared with the vectors from the original images considered as the ground-truth, for the image pairs on January 1 and 2. The histograms of the probability distribution functions appear on the bottom and right part of the plot, for the \( x \)- and \( y \)-axis drifts, respectively. ............................... 88
3.14 Tracking comparisons of sea ice fragments around the Beaufort Sea spanning from Feb. 17 to Mar. 18 in year 2013. The rows represent the tracking results on the ground-truth images (Orig.), the low-resolution input images (NN), and the super-resolved images (SR), respectively, for 4 indicative dates. Three ice fragment instances are tracked in the sequential frames as marked in bounding boxes with different colors. Context tracker [23] loses track of the tracking instances in low-resolution instances but keeps a stable performance over the super-resolved frames. .......................... 90

3.15 Success plots of the three-object tracking with the NN and SR Suomi NPP images compared with the results from the original images, for different overlap thresholds. The Area Under Curve (AUC) values are also provided. ....... 91

4.1 Enhancement results of “cone” (partial) (×4) in dataset [99]. Left column presents the low-resolution depth map and the intensity image and the right column shows the corresponding high-resolution results generated by the proposed enhancement framework. The low-resolution instances are upsampled by nearest-neighbor interpolation for a better illustration. ............... 96

4.2 Schematic flowchart of the proposed depth completion process. Pixel-level features are extracted based on the intensity image. The extracted features serve as the inputs to a SVM classifier which is learned via the training data labeled by the known values in the input depth map. A guided completion is then performed over the input depth map with the assistance of the label map generated by the SVM model and the aligned RGB image. ............... 98
4.3  Schematic pipeline of the super-resolution framework. To ensure a targeted learning, the input feature space is modeled with GMM. Within each Gaussian component, a regression model is learned. Given a low-resolution instance, the corresponding high-resolution instance is constructed through one of the regression models. ................................................................. 100

4.4  Resolution enhancement results of “Baby3”, “Midd1”, “Moebius”, and “Cone” (×4). From left to right, the columns represent low-resolution input depth maps, generated high-resolution depth maps, low-resolution intensity images, and generated high-resolution intensity outputs. For a better presentation, the input instances are upsampled through nearest-neighbor interpolation. ............................................................... 105

4.5  Resolution enhancement results in RGBD Scenes dataset v2 [64] (×4). From left to right, the columns represent low-resolution input depth maps, generated high-resolution depth maps, low-resolution intensity images, and generated high-resolution intensity outputs. For a better presentation, the input instances are upsampled through nearest-neighbor interpolation. ............................................................... 106

4.6  Super-resolution results of depth maps “bull” and “sawtooth” (×4). From left to right, the columns represent the ground-truth, results generated utilizing nearest-neighbor interpolation, patch-based [4], and the proposed framework. Only partial of the original depth maps are shown for a clearer presentation. 106

4.7  Super-resolution results of RGB images “bull” and “sawtooth” (×4). From left to right, the columns represent the results generated through nearest-neighbor interpolation, bicubic interpolation, ScSR [141], GR [120], and the proposed system. Only partial of the images are shown for a clearer presentation. 107
5.1 Sample images and descriptions in FlickrNYC dataset. (a) Examples from data_s in which images are with short descriptions. (b) Examples from data_l in which images are with long descriptions. 

5.2 Example of the description generated by the proposed sg-LSTM image captioning framework compared with the result output by the traditional multimodal RNN. Both frameworks are trained with FlickrNYC – our newly proposed dataset.

5.3 Systematic flowchart of the proposed sg-LSTM captioning framework. (a) Basic multimodal LSTM (m-LSTM) captioning framework trained on a subset of the FlickrNYC dataset with short descriptions (i.e., data_s). \( w_t \) denotes the \( t \)-th word in a sentence with words ranging from \( w_1 \) to \( w_T \). A start sign \( w_{start} \) and an end sign \( w_{end} \) are added to all training instances in both data_s and data_l. (b) Guiding text feature (GTF) extraction: to extract the text feature for self-guiding, we generate the descriptions utilizing m-LSTM followed by a sentence vectorizer. (c) Illustration of the sg-LSTM architecture. Compared with m-LSTM, an additional textual feature is fed to the multimodal block which encodes the language information connected to the image content.

5.4 Examples of the guiding text (marked in red) generated by m-LSTM compared with the original descriptions (marked in blue) provided by the Flickr users. The guiding text provides supplementary information that is strongly related to the image content.
5.5 Descriptions generated by the proposed framework (marked in red) compared with m-RNN [77], m-LSTM-long (m-LSTM trained on data), m-LSTM-full (m-LSTM trained on all training data) and the original descriptions (marked in blue) provided by the Flickr users. The guiding texts are also provided. To help with the evaluation, the ground-truth locations are marked in each image (usage of different colors is for the best contrast). . . . . . . . . . . . . 125
Chapter 1

Introduction

Vision research is generally divided into three subfields: low-level vision, middle-level vision, and high-level vision. Low-level vision focuses on the extraction of image properties from the retinal image, and is often based on low-level image processing. Middle-level vision emphasizes the integration of image properties into perceptual organizations. High-level vision is more related to the everyday functionality of perceptual organizations.

Among various vision related tasks, image enhancement aims to modify images to achieve a better perception for human visual system or a more suitable representation for further analysis in a variety of applications such as medical imaging, remote sensing, and video surveillance. It falls into low-level vision and has drawn increasing attention in improving image quality or interpretability. Based on the different attributes of the provided input images, tasks vary, for example, noise removal, deblurring, resolution enhancement, prediction of missing pixels, etc. The latter two are usually referred to as image super-resolution and image inpainting (or completion).

On the other hand, semantic image understanding belongs to high-level vision and is related to the analysis and interpretation of visual information. High-level objects, scene types, and their relations are studied to help with specific tasks such as object detection,
scene classification, image segmentation, and image captioning.

In this dissertation, three tasks, i.e., image super-resolution, image completion, and image captioning, are investigated spanning from low-level to high-level vision. Specifically, single image super-resolution and completion are studied in which only one single image is provided as input. Therein, ‘internal statistics’ refers to the statistics within the input image and across scales — such as raw pixel values or features extracted from the image — without the usage of external images or datasets. For image captioning, uncontrolled real-world setting is adopted in this study, i.e., the image-description training dataset utilized is collected directly from the web. In this topic, ‘internal statistics’ refers to the usage of only the noisy training dataset without other descriptions generated by human experts. In this Chapter, background information and related work of the three aforementioned tasks are presented.

1.1 Image Super-Resolution

Image super-resolution is to predict a fine-resolution image from one or multiple coarse-resolution images. It is one of the most fundamental image operations in image editing software. Recently, digital cameras are able to produce images with increasingly higher resolution. However, there still exist low-resolution images as well as low-grade sensors, e.g., in surveillance systems and mobile devices. Therefore, image resizing plays a crucial role in a variety of applications such as desktop publishing, movie restoration, and object tracking in satellite images.

In most of the literatures, the relation between a high-resolution image $I_H$ and a low-resolution image $I_L$ is simplified into a blurring and downsampling process:

$$I_L = (I_H \ast G_b) \downarrow_s$$

(1.1)

where $G_b$ represents the blur kernel, $\downarrow$ stands for the downsampling process and $s$ is the
CHAPTER 1. INTRODUCTION

scaling factor.

Image super-resolution itself is a numerically ill-posed problem. The upsampling process involves determining far more pixel intensities than the values given. This makes image super-resolution a particularly challenging problem and relies on additional information or assumptions to regularize the ambiguity and finalize the output image among all the possible candidate solutions. Even with different kinds of image priors, upsampled images usually lack small-scale texture-related features and moreover, sharp edges become blurry, original pixels grids remain noticeable, and in some cases, ringing appears in the vicinity of sudden transitions in intensity.

Multi-frame super-resolution methods [12, 30, 32, 39, 95, 105] make use of the presence of aliasing in multiple images of the same scene to produce one high resolution image. The basic idea behind this is to combine the non-redundant information contained in multiple low-resolution frames to generate a high-resolution image. For single image super-resolution, only one input image is available. In this dissertation, we focus on single image super-resolution.

Approaches addressing the single image super-resolution problem can be categorized as interpolation-based, reconstruction-based, and learning-based. Interpolation-based approaches have their roots in sampling theory and are commonly found in commercial software due to its efficiency. The produced results suffer from various edge-related visual artifacts such as ringing, blurring, aliasing, and jaggies. Reconstruction-based methods, also referred to as edge-directed methods [115], enforce some prior knowledge (often designed to generate sharp edges) on the upscaled high-resolution image. Performance of the reconstruction-based methods depends on the prior applied but often fails in synthesizing rich textured areas. Recently, learning-based techniques are very popular. This group of methods attempt to capture the co-occurrence prior between low-resolution and high-resolution images. Learning-based approaches can be further divided into statistics-based learning methods, and example-based learning methods.
CHAPTER 1. INTRODUCTION

1.1.1 Interpolation-based Super-Resolution

Linear interpolation is a classic and the simplest approach to predict intermediate pixel values. This method is usually implemented using data invariant linear filtering, such as nearest-neighbor, bilinear, bicubic, Hann, Hamming, and Lanczos interpolation kernels.

In [117], Thevenaz et al. presented a detailed survey of interpolation and resampling techniques. Interpolation is the model-based recovery of continuous data from discrete data within a known range of abscissa. Interpolation is based on the assumption that the underlying data is continuously defined. And different interpolation methods are proposed for image resizing.

Nearest-neighbor interpolation is the simplest among the listed interpolation methods and is made of a square pulse [117]. The unknown pixel value is assigned to be the most nearby translated pixel value. It requires the least processing time because it only considers the pixel closest to the interpolated point. However, the interpolated results look discontinuous.

Bilinear interpolation extends the neighborhood of the unknown pixel to a $2 \times 2$ area. Then the interpolated value is calculated through a weighted average of the 4 pixels in the neighborhood. Compared with nearest-neighbor interpolation, bilinear interpolation results are much smoother and more natural.

In contrast to bilinear interpolation, bicubic interpolation extends the neighborhood of the unknown pixel further to a $4 \times 4$ area with 16 pixels involved. Similarly, the interpolated value is calculated through combining the 16 pixels weightedly where closer pixels are given higher weights. Bicubic interpolation generates more visually favorable images compared with nearest-neighbor and bilinear interpolation. It is widely used in image editing software due to its stable performance.

Apodization is defined as the multiplication of a sinc function by some window [117]. A window is a mathematical function that is zero outside of some interval. Hann window
and Hamming window are two commonly used windows for interpolation. Hann window is defined as:

$$w_{Han}(n) = 0.5(1 - \cos\left(\frac{2\pi n}{N_t - 1}\right))$$  \hspace{1cm} (1.2)

where $N_t$ represents the width of a discrete time period and $0 \leq n \leq N_t - 1$.

Hamming window is defined as:

$$w_{Ham}(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N_t - 1}\right)$$  \hspace{1cm} (1.3)

Lanczos window is a sinc function windowed by the central hump of a dilated sinc function and is also often used to resize or rotate a digital image. It is defined as:

$$w_{Lanc}(n) = \text{sinc}\left(\frac{2n}{N_t - 1} - 1\right)$$  \hspace{1cm} (1.4)

Interpolation-based super-resolution approaches are based on the “smoothness” assumption which assume that image data is either spatially smooth or band-limited. The assumption is not the case in natural images. Natural images contain strong discontinuities, such as edges, corners, and high-frequency textured regions. Therefore, using interpolation kernels which are designed for spatially smooth or band-limited signals will result in visual artifacts such as ringing, aliasing, blocking, and blurring.

In order to suppress artifacts and restore sharper edges, more sophisticated interpolation methods \cite{72, 109} were proposed where interpolation weights adapt locally to the image content.

As stated above, bilinear interpolation uses all 4 pixels in the $2 \times 2$ neighborhood for interpolation. Unlike bilinear interpolation, in \cite{109}, interpolation weights are adjusted locally by choosing three out of the four nearest pixels to reduce the number of variables that
are averaged. Interpolator is tuned to match edges. By doing this, although edge blurring normally shown in classical interpolation methods is avoided, a noticeable block-like effect is observed in the generated results.

This artifact is avoided in [72] where local covariance coefficients from a low-resolution image are estimated and then based on the geometric duality between the low-resolution and the high-resolution covariance, those estimated coefficients are used to adapt the interpolation at a higher resolution image. The generated results have smooth curves and reduced jaggies. However, edges are still over-smoothed since the interpolation weights are estimated from flat regions as well.

1.1.2 Reconstruction-based Super-Resolution

Reconstruction-based image super-resolution approaches tend to estimate the target high-resolution image by enforcing some statistical priors during the upscaling process. Generally, these approaches also require the produced high-resolution image to be consistent with the input low-resolution image through strategies such as back-projection. Performance varies as different priors apply. The enforced priors are typically designed to reduce edge artifacts. Therefore, this group of approaches is also referred to as edge-directed super-resolution methods [115].

In [2], Aly and Dubois incorporated a total-variation regularizer into the object function. Since total-variation regularization discourages oscillatory isophotes, through minimizing the total variation functional, the low-resolution image is upsampled with smooth curves. However, there are still noticeable artifacts along the edges.

In [103], Shan et al. minimized a similar metric using a sophisticated feedback-control framework that keeps the output image consistent with the input image when downscaling it to the input resolution. The non-blind deconvolution process follows the structure proposed in [102] where global distribution of gradients is used for regularization during the
CHAPTER 1. INTRODUCTION

deconvolution.

The main drawback of edge-directed super-resolution approaches is their focus on preserving edges while leaving relatively “smooth” regions untouched. In [115], Tai et al. aimed to construct edges and recover image details. In their proposed framework, user guidance is needed to provide a single exemplar image to supply the missing details.

1.1.3 Learning-based Super-Resolution

Recently, learning-based super-resolution methods have been very popular due to the development of machine learning techniques and the availability of large natural image datasets. Among the numerous different approaches, learning-based solutions can be roughly divided into two categories: statistics-based learning and example-based learning. The former group tries to learn statistics from natural images and fit analytical models to describe various image features that show statistical dependency at different scales. The latter uses a patch- or feature-based approach to learn the relationship between local image details in low-resolution and high-resolution versions of the same scene. Example-based super-resolution can be further divided into subclasses based on different criteria, for example, patch-based or feature-based, whether external statistics are involved or not, etc.

Statistics-based Learning

Gradient statistics has been a popular learning prior for statistics-based super-resolution methods. It has been shown in [50] that natural image gradients generally follow a heavy-tailed distribution, indicating that most pixels have small gradient magnitudes whereas salient structures are present. Gradient density distributions can be used to formulate constraints during the image super-resolution process.

In [31], Fattal proposed an image super-resolution method which is based on the statistical edge dependency relating certain edge features of two different resolutions, which is
generically exhibited by real-world images. In addition to the edge dependency, the intensities are required to be conserved. Moreover, the output image must be identical to the input image when downsampled to the original resolution. For each pixel $x$ in the low-resolution image, three edge features extracted in [31] are:

- The total change in intensity across the nearest edge $m(x)$.
- The distance from the closest edge $d(x)$.
- The closest edge’s spatial scattering $s(x)$.

The high-resolution image is generated by constructing its gradient field rather than determining its pixel intensities directly.

Similarly, in [111], Sun et al. developed edge-specific priors to model the mapping of edges from coarse resolution to fine resolution to reconstruct sharp high-resolution edges. Different from [31], in [111], a fully analytical prior for the reconstructed edge profile is used.

These methods are usually considerably faster than example-based approaches, have stable performance and are successful in reproducing the sharpness of salient edges. However, due to their focus on edges or large scale features, they do not show improvement at small-scale textured and cluttered regions of the image. Moreover, a few parameters are far too insufficient to handle the various complex cases within a natural image. For the complicated textured regions, the generated results appear unrealistic as they are made of generic edges that often separate color plateaus.

Yang and Yang [139] proposed a divide-and-conquer approach to learn statistical priors directly from exemplar patches using a group of simple functions. The input space of low-resolution source images is divided and the mapping between low-resolution and high-resolution patches of each subspace is modeled by a linear function. The generated results have better performance in textured areas compared with the statistics-based methods mentioned above. However, it still suffers from visual artifacts such as blurring or over-smoothing.
Example-based Learning

Freeman et al. \cite{34, 35} proposed the example-based image upsampling method to learn the relationship between high-resolution images and their corresponding low-resolution ones through an external natural image dataset. The input image is subdivided into overlapping patches, which together form a Markov Random Field (MRF) framework. By searching for nearest neighbors in a low-resolution patch pool, a number of corresponding high-resolution candidates can be retrieved. This results in an MRF with a number of high-resolution candidate patches for each node. After associating a data cost to each candidate and a continuity cost for neighboring candidates, the MRF can be solved by using techniques such as belief propagation or graph cuts. To be more specific, a database of example patches are decomposed into their low-frequency band and the residual higher frequency band. The input image is interpolated to a higher resolution using analytic interpolation and the missing high-frequency band is then predicted from the patches. The matching is performed according to the low-frequency component of the example patches. By this means, an analytically interpolated image is enhanced by adding high-frequency patches from a non-parametric set of examples relating low and high resolutions.

This approach is capable of producing plausible fine details across the image, both at object edge and in fine textured regions. However, with lack of relevance between some test images and a universal training data set, the generated results are noisy with irregularities along curved edges. And since during the upsampling process, patch-wise comparisons are involved for each input patch and the whole dataset, the computational cost is very high. The performance may be improved with a larger database. The prospect of using a larger dataset for performance improvement is prohibitively time consuming, and matching small low-resolution patches is very limited in terms of distinguishing the proper examples. Approximate nearest-neighbor search offers a limited solution but also introduces its own
errors.

Many more example-based learning methods were proposed after [34, 35], we further divide them into two subclasses depending on whether they depend on external statistics or not.

*External Example-based Learning:*
Sun *et al.* extended [35] in [112] by using primal sketch priors to enhance blurred edges, ridges, and corners. In [116], Tappen *et al.* proposed a degradation model that allows a small number of discrete states to represent the large number of possible image values at each local image patch. Belief propagation is then utilized to find the best regressor to use at each point. The output is required to be consistent with the input. HaCohen *et al.* [43] refined the patch-based image model of [34] and interpreted the image as a tiling of distinct textures, each of which is matched to an example in a pre-defined database of relevant textures. The matching is done over the entire segments rather than the patches. User guidance is also needed in this approach. Similarly, authors in [113] also used segment exemplars for upsampling textures based on the observation that the consistency of high-resolution textures is better enforced by matching large segments than small independent patches.

As mentioned above, example-based learning with external dataset can be time consuming since the dataset used is usually large. Several methods have been proposed to overcome this high computational complexity, most notably neighbor embedding and sparse coding approaches according to [120].

Neighbor embedding super-resolution methods do not always explicitly focus on lowering computational complexity, but because their inherent interpolation of the patch subspace they can be used to lower the number of image patch exemplars needed, thus reducing the execution time. Neighbor embedding approaches assume that small image patches from a low-resolution image and its high-resolution counterpart form low-dimensional nonlinear
manifolds with similar local geometry.

Chang et al. [16] proposed a super-resolution method using the manifold learning method, i.e., Locally Linear Embedding (LLE). It assumes that each sample and its neighbors lie on or near a locally linear patch of the manifold when enough samples are available. The LLE algorithm is incorporated to address both ambiguity and dataset adequateness problems by integrating high-resolution exemplar patches with weights computed from their low-resolution patches. A set of $K$ nearest neighbors for each input patch is searched in the low-resolution feature space. Corresponding $K$ weights are computed for reconstructing the low-resolution patch by finding a constrained least squares solution, and eventually create a high-resolution patch by applying these weights in high-resolution feature space.

Recently, Bevilacqua et al. [9] also proposed another neighbor embedding method called Nonnegative Neighbor Embedding approach which is based on the assumption that the local nonnegative-least-square-decomposition weights over the local neighborhood in low-resolution space also hold for the corresponding neighborhood in high-resolution space. The use of a fixed number $K$ neighbors for reconstruction often results in blurring effect, due to over- or under- fitting.

Sparse coding approaches were firstly proposed by Yang et al. [141, 142] in learning a compact dictionary based on sparse signal representation. The low-resolution image is viewed as the downsampled version of a high-resolution image, whose patches are assumed to have a sparse representation with respect to an over-complete dictionary of prototype signal-atoms. This allows the possibility for adaptively choosing the most relevant reconstruction neighbors based on sparse coding, avoiding over- or under- fitting. Low-resolution patches are sparsely reconstructed using the following formulation:

$$\min_{\alpha} \|FD_l\alpha - Fy\|_2^2 + \lambda\|\alpha\|_0 \quad (1.5)$$
where $F$ is a feature extraction operator, $D_l$ is the learned low-resolution dictionary, $\alpha$ is the sparse representation, $y$ is the low-resolution input patch, and $\lambda$ is a weighting factor. Since the $l_0$-norm constraint leads to a NP-hard problem, in practice, the cost function is often relaxed to an $l_1$-norm constraint. Sparse dictionaries are jointly learned for low- and high-resolution image patches where low-resolution patches have the same sparse representation as their corresponding high-resolution patches. Eq. 1.5 can be easily reformulated to allow joint learning of both high-resolution dictionary and low-resolution dictionary. The resulting dictionary has a fixed size and is able to learn from numerous training patches while avoiding long processing times due to a growing size.

Zeyde et al. [145] built upon this framework and added a few modifications to improve the execution speed even more. The modifications include different training approaches for the dictionary pair, dimension reduction through PCA and Orthogonal Matching Pursuit for the sparse coding.

Timofte et al. [120] attempted to combine the benefits of both neighbor embedding and sparse coding. It starts with a learned sparse dictionary. Then for each dictionary atom, the nearest neighbors based on the correlation between the dictionary atoms are found. Afterwards, a separate projection matrix is computed for each dictionary atom based on its own neighborhood. The super-resolution problem is then solved by calculating, for each input patch feature, its nearest neighbor atom in the dictionary, followed by the mapping to high-resolution space using the stored projection matrix calculated offline.

Zhu et al. [147] proposed a deformable patches-based method which leads to a more “expressive” dictionary without increasing the size of the dictionary. By the concept of deformation, a patch is not regarded as a fixed vector but a flexible deformation flow. Via deformable patches, the dictionary can cover more patterns that do not appear, thus becoming more expressive.
Internal Example-based Learning:

Internal example-based image super-resolution methods have been proposed based on the observation that for small image patches in a natural image, self-similarities exist within the image itself and across different scales as illustrated in Fig. 1.1.

Glasner et al. [38] proposed a patch searching scheme based on a patch pool formed with internal patches collected through a pyramid structure with only the input image at different resolutions. The proposed framework combines the power of both the classical multi-frame super-resolution and the example-based super-resolution without any additional information. Recurrence of patches within the same image scale forms the basis for applying the classical super-resolution constraints. Recurrence of patches across coarser image scales implicitly provides examples of low-/high-resolution pairs of patches, thus giving rise to example-based super-resolution from a single image without any external database or any prior examples. The algorithm is able to produce visually pleasing results both in edges and fine textured regions.

Later, in [148], Zontak and Irani have demonstrated the effectiveness of internal statistics both in “Expressiveness” (how similar between a small patch and its most similar patches found internally or externally) and “Predictive Power” (how well can the found similar patches be used in image restoration tasks given a prediction model). For super-resolution task, it usually requires an external image dataset with hundreds of natural images to obtain similar outputs as compared with the results which use only internal statistics. Moreover, with the use of a large external dataset, the computational cost will increase dramatically.

Although external information is not used in [38], the algorithm is still computational demanding due to the pair-wise comparison for each patch with the self-formed patch pool. In order to increase the execution speed, Freedman and Fattal [33] proposed a method which follows a local self-similarity assumption and extracts patches from extremely localized regions in the input image. By keeping the scaling factor small, the local-self similarity
Figure 1.1: Patch recurrence within and across scales of a single image [38]. Source patches in \( I \) are found in different locations and in other scales of \( I \) (solid-marked squares). The high-resolution corresponding parent patches (dashed-marked squares) provide an indication of what the unknown high-resolution correspondences of the source patches might look like.
holds well. The small scalings are implemented using dedicated non-dyadic filter banks. The filters are nearly biothogonal and hence produce high-resolution images that are highly consistent with the input image without solving implicit back-projection equations. Although compared with [38], [33] is much faster, the results generated have visual artifacts of being “over-sharped” along the edges.

Similarly, Yang et al. [140] proposed a fast super-resolution method based on in-place example regression. The method refines the local self-similarity by adopting in-place self-similarity and proves that a patch in the upper scale has good matches around its origin location in the lower scale image. Based on the in-place examples, a first-order approximation of the nonlinear mapping function from low-resolution to high-resolution image patches is learned. It combines the learning from external statistics and internal statistics.

Blur kernel $G_b$ indicated in Eq. 1.1 is commonly set to be a Gaussian kernel, a bicubic kernel, or the Point Spread Function (PSF) in most super-resolution algorithms. Michaeli and Irani [85] have demonstrated that neither case mentioned above is true in practice. A framework for blind super-resolution was proposed in [85] to recover the correct super-resolution blur kernel. By using the correct blur kernel, both the external-based learning (demonstrated with [145]) and the internal-based learning (demonstrated with [38]) experience performance improvements.

### 1.2 Image Completion

Among the various situations of image enhancement, there are complicated circumstances where part of the input low-resolution image is undesired or unavailable, e.g., an overexposed spot, an unwanted shadow or scratch within a digital photo, or a satellite image partly covered by cloud. For these cases, in addition to the enhancement of image resolution, image completion techniques are also needed to reconstruct more visually plausible result.
The underlying goal for image completion is to predict the unknown pixel values within an image. It is also a numerically ill-posed problem and relies on additional assumptions or priors to finalize the output.

Different from internal example-based image super-resolution which incorporates self-similarities across different scales, example-based image completion approaches [18, 60, 84, 128, 136] deal with patch recurrence within the same resolution. They replace incomplete patches with similar ones in the known region and have been successful in replicating visually plausible background textures. Moreover, super-resolution techniques have already been utilized to assistant image completion. Le Meur and Guillemot [84] proposed a super-resolution-aided completion method which firstly performed completion in a coarse-version of the input image followed by a super-resolution process as to retain the original resolution.

Different from RGB images, depth maps are convenient in representing and storing the distance information of the objects’ surfaces given a viewpoint. They can be easily obtained through 3D imaging hardware such as time-of-flight (TOF) cameras and cost-effective consumer RGB-D cameras (e.g., Microsoft Kinect camera). Quality of the captured depth maps are crucial in their relevant applications, e.g., reliable 3D reconstruction, accurate human pose recognition, proper semantic scene analysis, and other geometry-related computer vision systems. However, due to the limitations of the depth sensors, depth maps suffer from low spatial resolution especially when the objects are far from the camera. Moreover, missing depth values exist due to the short distance between the object and the depth camera, disparity between the projector and the sensor, or poor reflection of the light patterns [105]. Under these circumstances, we reply on computer vision algorithms to enhance the quality of the depth maps.

To obtain high-quality depth maps, missing values in the input depth map need to be filled using image completion techniques. Image completion process aims at predicting the missing pixel values with the known regions and to successfully replicate visually plausible
background textures. In the depth domain, other than generating visually plausible results, the predicted depth values should be accurate in a manner that is consistent with the registered intensity images if available. Shen et al. [105] proposed a probabilistic model to capture various types of uncertainties in the depth measurement. Depth layers are utilized to achieve a depth correction and completion process where the layer labels are obtained through solving a maximum-a-posteriori estimation problem. In [56], with the assistance of an aligned high-resolution RGB image, an algorithm for simultaneous performing depth map super-resolution and completion is presented.

1.3 Image Captioning

In the recent popularized language-vision community, image captioning has been an important task. It involves generating a textual description that describes an image by analyzing its visual content. Automatic image captioning is able to assist solving computer vision challenges including image retrieval, image understanding, object recognition, navigation for the blind, and many others.

Although image captioning is a natural task for human beings, it remains challenging from a computer vision point of view especially due to the fact that the task itself is ambiguous. There are countless ways to describe one input image, from high-level descriptions to explanations in details, while all be semantically correct. The fundamental cause is that in principle, descriptions of an image can talk about any visual aspects in it varying from object attributes to scene features, or even refer to objects that are not depicted and the hidden interaction or connection that requires common sense knowledge to analyze [7].

In general, image captioning is a data-driven task. Descriptions for query images are normally defined by the training data. Therefore, it is not uncommon to see the birth of a new dataset for a new task. Recently, Amazon Mechanical Turk (AMT) is involved in
more and more dataset description generation process. Different sets of descriptions may be generated depending on the instructions provided to fit a specific captioning task. Since it is an expensive process, majority of the image captioning frameworks focus on exploring existing datasets which tend to provide a sentence description embedded with the objects, attributes, and the reactions with the scene in the image. Some other frameworks tackle the problem from a different angle, such as unambiguous descriptions [76], image stream descriptions [91], etc.

1.3.1 Methods

Based on the underlying models utilized, recent image captioning frameworks can be classified into three categories. The first group of approaches casts the problem as a retrieval problem in which description of a test image is generated by searching for similar images in a database. This group of models employs the visual space to measure the similarity during image search. Descriptions of these similar images are transferred to obtain the target description. In [137], image features are represented as the activations of the last layer of the Visual Geometry Group convolutional neural network (VGG-CNN) [108] trained on ImageNet [21]. The description of the query image is represented as a weighted average of the distributed representations of the retrieved descriptions. Different from [137], Devlin et al. [22] employed the n-gram overlap F-score between the descriptions to measure the description similarity.

The second group of methods adopts pre-defined sentence templates to generate image descriptions. The missing components in the sentence structures are filled based on image understanding of the objects, attributes, and the correlations between objects and the scene. Elliott and Keller [26] proposed a sentence generation model which parses a query image into a visual dependency representation (VDR) which then traversed to fill the missing slots in the templates. More linguistically sophisticated approaches [61, 86, 89] were proposed to
CHAPTER 1. INTRODUCTION

tackle the sentence generation.

The third group of approaches integrates image understanding and natural language generation into a unified pipeline. In general, image content in terms of objects, actions, and attributes is represented based on a set of visual features. Later, this content information is utilized to drive a language generation system, e.g., a recurrent neural network (RNN), to output the image descriptions. Some frameworks model image and text jointly into a multimodal space where later the joint representation space is used to perform cross-modal retrieval based on a query image. Karpathy and Li [54] presented an alignment model which uses a structured object to align the two modalities (i.e., CNN over image regions and bidirectional RNN over sentences) through a multimodal embedding. In [58], an encoder-decoder framework is presented utilizing a joint multimodal space in which the long short-term memory (LSTM) is a big success. Another represented work in this category is the m-RNN model [77] in which a multimodal component is introduced to explicitly connect the language model and the vision model by a one-layer representation.

With image captioning being a thriving topic, it is driven by the technical trials and improvements in both computer vision and natural language processing, and also importantly, the availability of relevant datasets. Other than the traditional image captioning task, efforts have been made to special captioning tasks. In [78], Mao et al. modified m-RNN to address the task of learning novel visual concepts. Authors in [44] incorporated unpaired image data with labelling and unpaired text data to address the concept limitations in the image-sentence paired dataset. ‘Referring expression’ is explored in [55, 76] to generate unambiguous descriptions. Park and Kim [91] presented a coherence recurrent convolutional network (CRCN) to describe an image stream in a storytelling manner utilizing blog data.
1.3.2 Datasets

Due to the rising interest in image captioning task, a number of datasets have been brought up varying in sizes, formats of the descriptions, and the collection process. One of the earliest benchmark datasets — Pascal1K [96] was proposed by Rashtchian et al. which consists of 1,000 images selected from Pascal 2008 object recognition dataset [29]. Each image is associated with five sentence descriptions generated by AMT.

Later, based on Pascal 2010 action recognition dataset, Elliott and Keller introduced the Visual and Linguistic Treebank (VLT2K) [26] with 2,424 images. AMT is again utilized with specific instructions to generate three, two-sentence descriptions for each image. Object annotation is available for a small subset of the images and VDRs are created manually for these images.

The Flickr8K [48] and Flickr30K [144] find their roots on images from Flickr. Although the images are collected based on user queries for specific objects or actions, the descriptions are generated in a manner similar to Pascal1K dataset where AMT workers provide five captions for each image. The original titles or descriptions from Flickr are not directly utilized to generate the captions in these two datasets. On the other hand, user-provided descriptions are employed in SBU1M [88] which contains approximately one million captioned images from Flickr. Strict filtering is applied that the downloaded image should contain at least one noun and one verb on predefined control lists.

The MS COCO dataset [74] is widely used recently for image captioning evaluation with 123,287 images accompanied by five descriptions per image. Extensions of MS COCO dataset are available to meet specific needs of various tasks, e.g., question answering [3], unambiguous descriptions [76], and text detection and recognition [125]. The Déjà image captions dataset [17] makes use of the existing web data without additional human efforts. It consists 4 million images with 180K unique captions where lemmatization and stop word
removal are employed to normalize the captions and create a corpus of near-identical texts.

Although various datasets have been collected recently, expensive human labelling under specific instructions or strict filtering is often required especially for large datasets. However, as mentioned, image descriptions should come in different degrees of abstraction, i.e., descriptions could be abstract as in several words or a short term, or as a detailed paragraph in a storytelling way.

Contributions of this dissertation are summarized as follows:

• An internal gradient similarity-based super-resolution framework is proposed and extended to handle image super-resolution and completion simultaneously.

• To benefit from both internal and external statistics, a hybrid example-based super-resolution algorithm is proposed. A test case application of this super-resolution method enhancing passive microwave derived sea ice motions is presented to link the algorithm design with a real-world application.

• An image quality enhancement system is proposed to improve the quality of paired low-resolution RGB image and depth map.

• A self-guiding multimodal LSTM framework is proposed to accomplish image captioning task based on uncontrolled real-world web data.

The rest of the dissertation is organized as follows: Chapter 2 introduces the proposed internal gradient similarity-based super-resolution and completion framework. Chapter 3 provides the detailed information of the second single image super-resolution framework, i.e., hybrid example-based super-resolution algorithm, and its application. In Chapter 4, we introduce the system which improves the quality of paired low-resolution RGB image and depth map where there are missing values in the input depth map. The self-guiding multimodal LSTM captioning framework is presented in Chapter 5. The conclusions are drawn in Chapter 6.
Chapter 2

Internal Example-based Image Enhancement

In this chapter\(^1\), we first introduce a single image super-resolution framework which is based on internal across-scale gradient similarity. The use of internal and external statistics in image super-resolution is investigated and extensive experimental results are presented.

Afterwards, in Section 2.2, the proposed framework is extended to handle complicated circumstances where part of the input low-resolution image is unavailable, i.e., to accomplish image super-resolution and completion simultaneously.

2.1 Internal Gradient Similarity-based Super-Resolution

2.1.1 Framework Details

In this subsection, we introduce the proposed image super-resolution method based on internal across-scale gradient similarity. Fig. 2.1 illustrates the schematic pipeline of the framework. Since human eyes are more sensitive to brightness changes than color changes, there-

\(^1\)This chapter was previously published in [131, 133].
Figure 2.1: Flowchart of the proposed across-scale gradient similarity-based super-resolution algorithm. After calculating the gradients of the input low-resolution image in horizontal and vertical directions (represented as $L_x$ and $L_y$), for each gradient patch, its top $k$ most similar patches are searched within the corresponding gradient patch pool. Patch pool $\mathcal{P}_x$ is composed of all gradient patches in the downsampled version of $L_x$ (represented as $LL_x$). $\mathcal{P}_y$ is built up in a similar manner utilizing $L_y$. After constructing the high-resolution gradients $H_x$ and $H_y$, the output image is restored based on them and the input image $L$.

Therefore, same as the majority of other super-resolution approaches, for a given input color image, the proposed algorithm is only performed in the luminance channel of the YUV color space while the other two channels are upsampled through bicubic interpolation.

Given a grayscale low-resolution image $L$, to upscale $L$ by a scaling factor of $s$, we first calculate the gradients $L_x$ and $L_y$ of $L$ in horizontal and vertical directions. Afterwards, based upon internal across-scale gradient similarity, $L_x$ and $L_y$ are upsampled individually by $s$ to obtain the gradients in high-resolution represented as $H_x$ and $H_y$, respectively; the second step is to reconstruct the target high-resolution image $H$ from $L$, $H_x$, and $H_y$ through optimizing a uniform cost function which incorporates the constraints in both image level and gradient level.
Upscaling Low-Resolution Gradients

In order to upscale the low-resolution gradient in horizontal direction by factor $s$, $L_x$ is first decomposed into a set of overlapping patches at size $a \times a$ ($a = 5$ in our implementation) with stride equals to 1. $LL_x$ is calculated by downsampling $L_x$ by $s$. A gradient patch pool $\varphi_x$ is constructed with all the patches with size $a \times a$ in $LL_x$. In order to form a more expressive patch pool, all the patches are normalized to have zero mean and uniform variance to better preserve the structural information.

Gradients of the natural images have been modeled by a heavy-tailed distribution. Therefore, generally for natural images, gradient patches form a sparse distribution. Majority of the gradient patches will be flat with small variances. For these patches, bicubic interpolation will be effective enough without compromising the final super-resolution performance.

For each patch $p$ in $L_x$, we calculate its variance and compare it with a pre-set threshold $\theta_v$. If the variance is smaller than $\theta_v$, $p$ is upscaled directly through bicubic interpolation; otherwise, after patch normalization, its top $k$ most similar patches are searched within gradient patch pool $\varphi_x$. The similarity between two patches is measured in their mean square error.

Within a pair of images representing the same scene but at different resolutions, given an instant patch in the coarse-resolution image, the corresponding patch in the high-resolution image is referred as its “parent” patch. After obtaining the top $k$ most similar patches within $\varphi_x$ for query patch $p$, their “parent” patches in $L_x$ are extracted, normalized to have zero mean and unit variance, and then combined weightedly. Patches that are more similar to the query patch are assigned with larger weights. The weights used to combine the $k$ patches are computed with:

$$w_i = \exp\left(-\frac{\sum_{j=1}^{k} M_i M_j}{R}\right)$$

(2.1)

Therein, $w_i$ represents the weight for patch $i$ during the combination, $M_i$ stands for the mean
square error between the query patch \( p \) and patch \( i \), \( R \) is a normalization factor to ensure the summation of the \( k \) weights equals to 1.

The combined patch is then adjusted according to the original mean and variance value of the input patch \( p \) and then "pasted" to \( H_x \) in position corresponding to patch \( p \) in \( L_x \). The overlapping area between adjacent patches is simply averaged. \( H_y \) is calculated in a similar manner utilizing \( L_y \) and \( \psi_y \).

**Reconstruct High-Resolution Image**

With \( L \), \( H_x \), and \( H_y \), the target high-resolution image \( H \) is reconstructed by optimizing the following cost function:

\[
H^* = \arg\min_H \{|(H \ast G) \downarrow - L|^2 + \lambda |\nabla H - \nabla H_D|^2\}
\]  

(2.2)

where \( \nabla H_D \) represents the calculated \( H_x \) and \( H_y \). \( \downarrow \) represents the downsampling operation. \( G \) stands for a Gaussian kernel. We set its standard variance \( \sigma \) related to the upsampling scale \( s \) same as in [111]: \( \sigma = 0.8 \) if \( s = 2 \); \( \sigma = 1.2 \) if \( s = 3 \); \( \sigma = 1.6 \) if \( s = 4 \).

We integrate constraints in both image level and gradient level into a single cost function as shown in Eq. 2.2: the first term ensures the consistency between the output high-resolution image and the input low-resolution image. It has been demonstrated in [25] that a global constraint in the fidelity between input and output images is critical in image super-resolution; the second term constrains the gradients of the reconstructed high-resolution image to be close to the calculated high-resolution gradients. The parameter \( \lambda \) controls the weight between these two terms. The cost function can be minimized through the gradient descent algorithm with:

\[
H^{t+1} = H^t - \delta \cdot ((H^t \ast G) \downarrow - L) \uparrow \ast G - \lambda \cdot (\nabla^2 H^t - \nabla^2 H_D))
\]  

(2.3)
where $H^t$ represents the output after the $t$-th iteration and $\delta$ indicates the step width.

### 2.1.2 Why Gradient Level Works Better?

In the proposed super-resolution framework, patch-based self-similarity is utilized. Different from the internal example-based super-resolution methods [33, 38, 140], which perform the upscaling algorithms in the image level, we instead calculate the high-resolution gradients based on internal gradient patch similarity. The target high-resolution image is then reconstructed based on the generated high-resolution gradients and the input low-resolution image. In this subsection, two questions are answered:

- It is easy and straightforward to calculate the gradients in horizontal and vertical directions given an image. How well can we estimate a high-resolution image from its corresponding low-resolution image and the high-resolution gradients?

- While the proposed algorithm can be utilized to upscale an image directly, why instead taking a detour to upsample the gradients first?

### Reconstruct Image from Gradients

The feasibility of recovering images from their gradients are investigated. The reconstruction process is based on Eqs. 2.2 and 2.3. We verify that the target high-resolution image can be well reconstructed from the input low-resolution image and the corresponding high-resolution gradients.

We conduct an experiment based on the Berkeley Segmentation Dataset (BSDS) [79]. For each high-resolution image $H_{GT}$, its gradients in horizontal and vertical directions are calculated and serve as $\nabla H_D$ (as shown in Eq. 2.2). The low-resolution image $L$ is obtained through downsampling $H_{GT}$ by the factor $s$. Then $H$ is restored iteratively according to Eq. 2.3. $H^0$ is initialized as the bicubic interpolated version of $L$. After the reconstruction,
we calculate the difference between the reconstructed image $H$ and the original ground-truth image $H_{GT}$.

Fig. 2.2(a) demonstrates the similarity of the original images and the reconstructed images. The experiments are performed over 100 natural images in database BSDS [79] at different scales $s$ and weights $\lambda$. We adopt three different criteria in measuring the similarity between the ground-truth images and the reconstructed ones. The pixel-wise averaged error is calculated as the absolute difference between the reconstructed image and the ground-truth image followed by a division of the total number of pixels within each image. For upscaling factors of 2 and 4, an increase in weight $\lambda$ consistently results in the improvement of the reconstruction performance measured in the increment of Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [127] and the decrement in the pixel-wise averaged error due to the absolute accuracy of $\nabla H_D$. However, in practice, it is impossible to restore the perfectly accurate high-resolution gradients from the low-resolution gradients. To better control the visual artifacts in the output high-resolution images, we set $\lambda$ as 0.2 in our final super-resolution experiments. More details and discussions can be found in Sec. 2.1.4.

The effectiveness of restoring an image from its gradients is further demonstrated in Fig. 2.2(b)-(g) with two specific examples. Here the scaling factor is 4 and the weight is set to 1. Fig. 2.2(b) is the ground-truth “child” image. The low-resolution image is generated by downsampling the ground-truth image by factor 4. Fig. 2.2(c) is the reconstructed image based on the ground-truth gradients. As seen clearly from the zoom-in regions, edges and textures are perfectly restored even for the eyelashes. Visually we cannot differentiate Fig. 2.2(b) from (c). Fig. 2.2(d) presents the pixel-wise differences between Fig. 2.2(b) and (c) which further proves that the ground-truth image and the reconstructed image are nearly identical. Fig. 2.2(e)-(g) provide another set of results over image “smile”. The ground-truth image is a little noisy. Still, the reconstructed image is almost the same as the original one. In both examples, as shown in Fig. 2.2(d) and (g), majority of the reconstruction error lies
Figure 2.2: Demonstration of the feasibility to restore the high-resolution image from accurate horizontal and vertical gradients and the corresponding low-resolution image using the proposed reconstruction framework. (a) Experiments over the ground-truth images and the reconstructed images based on 100 natural images in database BSDS [79] at different scales and weights ($\lambda$). Left: pixel-wise averaged error; Middle: PSNR (dB); Right: SSIM. (b)-(g) illustrate the reconstruction results on images “child” and “smile” along with the ground-truth images and the corresponding difference maps. Error mainly exists along image boundaries.
in the boundaries of the images due to the fact that we do not have adequate information in those regions. We observe this similar pattern for all the rest images involved in the experiment. Extremely good PSNR and SSIM numbers are listed to illustrate that a near-perfect high-quality image can be well reconstructed from accurate gradients and the corresponding low-resolution image.

**Gradient Similarity vs. Image Similarity**

Accurate reconstruction of edges and textures are critical to super-resolution since they are the most perceptually essential features in a natural image. However, it is difficult to automatically hallucinate both edges and textures within one framework due to the different characteristics revealed by these two features. In this dissertation, the edges which provide structural information of the objects in the images are referred as structural edges. Therefore, in our proposed method, we aim to create sharp structural edges and visually plausible fine textures.

As mentioned, a key element in super-resolution is that the output high-resolution image should be consistent with the input low-resolution image. It is demonstrated in [25] that ensuring the consistency during high-resolution image reconstruction is at least as important as the usage of a proper image prior. Internal example-based super-resolution approaches ensure a local consistency between low-resolution patch instance and the corresponding high-resolution patch. A global fidelity constraint is often ignored in the upsampling scheme. Although back-projection is commonly utilized, it only provides a limited solution. In the proposed framework, both local and global fidelities are ensured in one uniform framework represented by two different terms in the cost function.

Traditional internal example-based super-resolution methods have been successful in synthesizing rich details. Compared with edge-directed approaches, this group of methods hallucinates more natural textures. However, it is difficult to control the artifacts introduced
CHAPTER 2. INTERNAL EXAMPLE-BASED IMAGE ENHANCEMENT

along structural edges especially under large scaling factors. On the other hand, gradients emphasize more on the intensity changes. Modeled by a marginal distribution, image gradient is often combined with $L_2$ norm or sparsity regularization and has been a success in a variety of image restoration tasks. As illustrated in [111], reconstruction from only the image level or only the gradient level introduces artifacts. Combining constraints in both levels provides better and more stable super-resolution performance. Therefore, the scheme that uses gradient patches combined with the proposed reconstruction framework is more robust compared with the traditional self-similarity approaches performed in the image level. However, gradient-based approaches are known to be noise sensitive. Under circumstances where the input images are very noisy, the proposed super-resolution framework could be combined with the denoising algorithms to enhance the overall image quality in a sequential manner.

To further demonstrate the effectiveness of the proposed gradient similarity-based super-resolution scheme, we reconstruct the high-resolution images from the same low-resolution input image by magnification factor of 4 using the proposed “search and paste” framework based on image similarity and gradient similarity, respectively. The two cases use the same set of parameters to upsample a low-resolution image or gradient. As demonstrated in Fig. 2.3, the results generated utilizing gradient similarity produce sharper edges and more natural details than those based on image similarity. For the two examples (shown in the top and middle rows in Fig. 2.3) with the ground-truth images available, results generated used the proposed gradient-based scheme are closer to the ground-truth images and the image-level results are over-blurred in structural edges and textures.

2.1.3 Internal vs. External Statistics

The use of external natural images from a large dataset to construct high-resolution images has raised many discussions recently. In this subsection, we investigate the contribution of
Figure 2.3: Comparisons of the super-resolution results ($\times 4$) in the image level and the gradient level (with reconstruction) respectively using the same framework. Three examples are presented and the ground-truth images are included for two cases shown in the top and middle rows. The ground-truth image is not available for the example in the bottom row. In all three cases, it is clearly demonstrated that the results produced in the gradient level reveal sharper edges and more natural textures.
internal and external statistics for gradient similarity-based super-resolution reconstruction. We define the internal statistics to be the image or gradient patches extracted from only the input image at different resolutions without using any external sample images or statistics; the external statistics represent image or gradient patches extracted from images of an external image dataset.

Zontak and Irani [148] have demonstrated the effectiveness of internal statistics both in “Expressiveness” (how similar between a small patch and its most similar patches found internally or externally) and “Predictive Power” (how well can the found similar patches be used in image restoration tasks given a prediction model). For super-resolution tasks, in order to achieve comparable results with methods adopt internal statistics, it usually requires a large external image dataset with hundreds or thousands of natural images. However, by the use of a large external image dataset, the computational cost will increase dramatically. Moreover, increasing the size of the external dataset makes the patch correspondences even more ambiguous [147].

We aim at exploring, with the presence of internal statistics, whether the external gradient statistics is helpful in boosting the performance of gradient similarity-based super-resolution. Since including a large external dataset will be infeasible in practice, our experiments in this section collect external statistics from a small dataset (with 5 or 10 high-quality images). Even though the number of external images used is limited, under a small patch size (i.e., $5 \times 5$), still hundreds of thousands of patch instances are collected.

We compare the reconstructed high-resolution images from the same low-resolution input image using only internal gradient statistics and both internal and external statistics. For both cases, the image super-resolution follows the same pipeline as introduced previously. The only difference lies in the formed gradient patch pools. Two different kinds of external statistics are evaluated: general external gradient statistics and class-specific external gradient statistics.
Table 2.1: Results of the percentage of images which have larger SSIM when general external statistics is introduced at three different input sizes.

<table>
<thead>
<tr>
<th>Input size</th>
<th>Percentage of images with larger SSIM with general external statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 × 40</td>
<td>85%</td>
</tr>
<tr>
<td>80 × 80</td>
<td>56%</td>
</tr>
<tr>
<td>160 × 160</td>
<td>33%</td>
</tr>
</tbody>
</table>

**General external gradient statistics**

External statistics is collected from 10 high-resolution natural images. We evaluate the contribution of the external gradient statistics in constructing 100 high-resolution images utilizing BSDS dataset [79]. All the images are rescaled to three sizes: 40 × 40, 80 × 80, and 160 × 160. The super-resolution task is to upsample these images by a scaling factor of 2 using only internal gradient statistics and by using both internal and external statistics. In both cases, the upscaling schemes follow the same pipeline as shown in Section 2.1.1. The only difference lies in the formation of the gradient patch pool. Only internal gradient patches are utilized to form the patch pool if internal gradient statistics is adopted. For the latter case, besides internal collected patches, gradient patches from the downsampled version of the 10 external images are also included in the patch pool. Since the original sizes of the 100 images are larger than 320 × 320, the ground-truth images are available for all three input sizes.

Evaluation of the super-resolution performance based on different statistics is measured with SSIM [127] between the generated high-resolution images and the ground-truth images. Here we assume that a larger SSIM indicates a better super-resolution performance if based on a uniformed pipeline.

Table 2.1 illustrates the percentage of images which have an increase in SSIM when ex-

---

2The images are downloaded from Flickr (www.flickr.com).
ternal statistics are introduced for each input size. For most of the input images with size $160 \times 160$, including external statistics will not boost the super-resolution performance measured in SSIM. On the contrary, 67% of the images have a larger SSIM without external statistics. An increase in the patch pool will definitely increase the “Expressiveness”. However, it will reduce the “Predictive Power” for the super-resolution task in general. As the size of the input image gets smaller, external statistics becomes more useful since the patch pool formed with only internal statistics will be very limited.

Generally, for super-resolution applications, the size of the input image is in hundreds by hundreds pixels, therefore, in our proposed super-resolution model, only internal statistics is adopted without the use of general external statistics.

**Class-specific external gradient statistics**

Zontak and Irani [148] have described the class-specific external database to be “extremely useful, even if small”. Here, we interpret class-specific external database as a set of high-resolution images which are highly similar to the target high-resolution image (images representing the same scene or sharing the same textures). We again evaluate the performance for image super-resolution task by using only internal statistics and both internal and class-specific external statistics.

The experiment is conducted with the UIUC Texture Dataset [66]. The database includes 25 texture classes, 40 samples for each class, all in grayscale (refer to Fig. 2.4 for examples of different texture patterns). We run the experiment on all the 25 classes. For each class, the first 5 images are utilized to collect the external gradient statistics; the remaining 35 images are downsampled by magnification factor of 2 and then serve as the input images for the evaluation. After the rescaling, the input low-resolution images all have the size $320 \times 240$.

Table 2.2 illustrates the percentage of images which have larger SSIM when class-specific external statistics is introduced. The results demonstrate that with the presence of small
Figure 2.4: Example images of all the 25 classes in the UIUC Texture Dataset [66] (sorted in descending order measured in “percentage of images with larger SSIM with class-specific external statistics”. Refer to Table 2.2 for more details.)
<table>
<thead>
<tr>
<th>Image class</th>
<th>Percentage of images with larger SSIM with class-specific external statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>bark1</td>
<td>74.29%</td>
</tr>
<tr>
<td>bark2</td>
<td><strong>100.00%</strong></td>
</tr>
<tr>
<td>bark3</td>
<td>45.71%</td>
</tr>
<tr>
<td>wood1</td>
<td>37.14%</td>
</tr>
<tr>
<td>wood2</td>
<td>22.86%</td>
</tr>
<tr>
<td>wood3</td>
<td>65.71%</td>
</tr>
<tr>
<td>water</td>
<td>34.29%</td>
</tr>
<tr>
<td>granite</td>
<td>80.00%</td>
</tr>
<tr>
<td>marble</td>
<td>14.29%</td>
</tr>
<tr>
<td>floor1</td>
<td>71.43%</td>
</tr>
<tr>
<td>floor2</td>
<td><strong>0.00%</strong></td>
</tr>
<tr>
<td>pebbles</td>
<td>82.86%</td>
</tr>
<tr>
<td>wall</td>
<td>77.14%</td>
</tr>
<tr>
<td>brick1</td>
<td>88.57%</td>
</tr>
<tr>
<td>brick2</td>
<td>74.29%</td>
</tr>
<tr>
<td>glass1</td>
<td>71.43%</td>
</tr>
<tr>
<td>glass2</td>
<td>57.14%</td>
</tr>
<tr>
<td>carpet1</td>
<td>82.86%</td>
</tr>
<tr>
<td>carpet2</td>
<td>80.00%</td>
</tr>
<tr>
<td>upholstery</td>
<td>91.43%</td>
</tr>
<tr>
<td>wallpaper</td>
<td>71.43%</td>
</tr>
<tr>
<td>fur</td>
<td>48.57%</td>
</tr>
<tr>
<td>knit</td>
<td>80.00%</td>
</tr>
<tr>
<td>corduroy</td>
<td>28.57%</td>
</tr>
<tr>
<td>plaid</td>
<td>91.43%</td>
</tr>
</tbody>
</table>

Table 2.2: Results of the percentage of images which have larger SSIM when class-specific external statistics is introduced at 25 different classes in the UIUC Texture Dataset [66].
scale class-specific external statistics, the super-resolution results vary for different classes. For class “bark2”, all images have better performance when external statistics is introduced. However, for class “floor2”, none of the images achieves better performance with external statistics. Fig. 2.4 presents the examples of all the 25 texture patterns from the UIUC texture database. The 25 texture patterns are sorted in descending order measured in “percentage of images with larger SSIM with class-specific external statistics” as presented in Table 2.2. There is no clear criterion to differentiate the classes which benefit from class-specific external statistics with those do not. As observed from Fig. 2.4, very roughly speaking, images with sparse or regular square patterns tend to benefit more when class-specific external statistics is introduced; on the other hand, external gradient statistics from dense irregular or parallel patterns degrade the super-resolution performance.

To sum up, our experiments indicate that external gradient statistics collected from a small dataset normally does not improve the performance of general super-resolution tasks. However, under certain circumstances, such as when the size of the input image is very small or the external dataset contains fine-resolution images very much similar to the target high-resolution image, external statistics might be helpful. The proposed super-resolution method adopts only the internal statistics but can be easily scaled to include external statistics if necessary.

2.1.4 Experimental Results

The proposed super-resolution method is evaluated with multiple images at different upsampling factors. The grayscale images are directly upscaled using the proposed algorithm. For the color images, as we mentioned previously that since human eyes are more sensitive to luminance changes, we only perform the proposed algorithm on the luminance channel in YUV color space while the rest two channels are upscaled through bicubic interpolation.
CHAPTER 2. INTERNAL EXAMPLE-BASED IMAGE ENHANCEMENT

Parameter Selection

Same as many existing state-of-the-art super-resolution methods, large upscaling factors require a coarse-to-fine scheme in this proposed model. However, instead of taking a relatively small scaling factor per-step, our method upsamples the image by a factor of 1.5 or 2 progressively, i.e., upscaling factor of 3 follows a $1.5 \times 2$ manner, upscaling factor of 4 takes steps $2 \times 2$, upscaling factor of 8 is calculated as $2 \times 2 \times 2$, and so forth.

Patch size $a$ during the gradient patch upsampling is set to be 5 and during the searching, the top 5 most similar patches are selected. Threshold $\theta_v$ in differentiating smooth patches from non-smooth patches is set to 10. To reconstruct the final output image from the upscaled gradients, the weight $\lambda$ is assigned to be 0.2. As mentioned earlier, although with the presence of the ground-truth gradients, a larger weight leads to a better reconstruction result, in practice, we do not have the gradients which are extremely accurate available. Output images generated by a relatively large $\lambda$ suffer from visual artifacts of over-sharped edges and unrealistic textures. Setting $\lambda$ to a large number will also increase the noise sensitivity of the proposed super-resolution framework. To find the proper $\lambda$, a set of high-resolution images are generated at scales 2 and 4 in datasets BSDS [79] and SET5 [9] utilizing different $\lambda$ ranging from 0.05 to 0.5 at a stride of 0.05. We observe that $\lambda$ between 0.1 and 0.2 gives fairly similar outputs with stable super-resolution performance. In Eq. 2.3, with a fixed step, a larger $\lambda$ will lead to a faster convergence. Therefore, we set $\lambda$ to 0.2 in our final experimental setting.

Visual Results

The proposed method is evaluated with a variety of natural images under different upscaling factors. We also compare our results with the state-of-the-art approaches [31, 33, 38, 103, 120, 139, 140]. Single image super-resolution is very challenging when the scaling factor is
large due to too much missing information. We need to estimate numerous unknown values based on a small amount of given pixels. In the coarse-to-fine scheme, error accumulates as the scaling factor increases. Fig. 2.5 demonstrates the feasibility of the proposed super-resolution method when the magnification factor is large (8 for the image “comic” and 16 for the image “butterfly”). The proposed algorithm is robust with visually pleasing super-resolution performance and is capable of reconstructing realistic high-frequency details.

We further compare our results with other approaches as illustrated from Figs. 2.6−2.9. Fig. 2.6 presents a set of super-resolution results on image “child” with an upscaling factor of 4. “Child” has been a popular test image in single image super-resolution papers. Although recent state-of-the-art approaches are able to generate clean edges along eyes, face, and lip contours, it is still very challenging to restore the knit textures in the hat region. As clearly illustrated in the zoom-in areas, our approach can produce more realistic hat textures with minimal artifacts compared with peer results. Moreover, in our generated high-resolution image, the contour between face and hat is more natural without being too “sharp” so that the hat is actually above the face, not vice versa. Yang et al. [140] is able to restore sharp and clean structural edges but tends to blur the hat area which contains complicated knit patterns. Timofte et al. [120] and Yang et al. [139] are capable of generating natural-looking outputs but there are noticeable artifacts along the face contour and within the eye areas. The hat region is also over-smoothed with blurry visual artifacts. In contrast, our algorithm successfully produces clean face, lip contours and reconstructs hat textures closest to the ground-truth image.

Fig. 2.7 provides another set of super-resolution results over image “chip” which is also a commonly used test image. The magnification factor is 4. The zoom-in areas clearly indicate that our method creates natural results of the characters and along the structural edges of the chip.

In Fig. 2.8, we compare our results with two dictionary-based approaches [142, 145]
Figure 2.5: Super-resolution of image “comic” (×8). (a) Input low-resolution image; (b) Generated high-resolution result. Super-resolution of image “butterfly” (×16). (c) Input low-resolution image; (d) Generated high-resolution result. Our method synthesizes fine details and restores clear edges. For a better presentation, the input image is upscaled to the target resolution utilizing nearest-neighbor interpolation.
Figure 2.6: Super-resolution of image “child” (×4). Our proposed method successfully reconstructs the knit textures in the hat region and maintains sharp and natural eyes, facial and lip contours so that the hat is actually above the face, not vice versa as observed in peer results.
CHAPTER 2. INTERNAL EXAMPLE-BASED IMAGE ENHANCEMENT

Figure 2.7: Super-resolution of image “chip” ($\times 4$). Our proposed method retains clean and sharp structural edges.
Figure 2.8: Super-resolution of images “baboon”, “Lenna”, and “pepper” (×3). In all three images, our proposed method generates results with clear edges and realistic textures closest to the ground-truth images.
Figure 2.9: Super-resolution of images “mushroom”, “flower”, “car”, “fish”, and “girl” (×4). Our proposed method generates results with both sharp edges and natural textures.

with the existence of the ground-truth images under magnification factor 3. The three example images contain different kinds of challenging cases for super-resolution including fur, whiskers, eyes, fabrics, feather, and shadows. As shown in the zoom-in areas, our produced results are vivid and realistic with more natural details compared with the peer methods. For example, in image “Lenna”, the edges in the hat fabrics generated by [142, 145] are over-smoothed and our result reconstructs patterns very much similar to the ground-truth image. Moreover, the reconstructed feather in the hat by our proposed method is more natural.
Table 2.3: IFC comparisons with peer super-resolution methods on SET5 [9].

More results are presented in Fig. 2.9 compared with Shan et al.\textsuperscript{3} [103], Sun et al. [111], and Yang et al. [139]. The ground-truth images all come from BSDS dataset. In the image “mushroom”, the results generated by [103, 111] tend to blur the patterns within the mushrooms while there are noticeable visual artifacts in [139]. Our result better synthesizes the complicated details without over-sharpening the edges. Our method also produces sharper edges with minimal artifacts in the “flower” image as illustrated in the zoom-in region. For the “fish” image, all the other three methods fail to recover the white line along the body contour of the fish. But our method nicely reconstructs this structural edge. Similarly, for the tire part in “car” and the cloth shown in “girl”, our results are clearer and more visually pleasing.

Quantitative Evaluation

The super-resolution performance is further evaluated quantitatively in this subsection. Recently, in [138], a variety of image quality metrics are evaluated by investigating the correlation with visual perception by human experts. Among the 8 criterions evaluated, Information Fidelity Criterion (IFC) [104] seems to be the most proper metric that could be utilized to perform comparison among different super-resolution frameworks. Therefore, we compare our results with [103] and [142] on dataset SET5 [9] (with 5 images, i.e., “baby”,

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
 & SET5 & Shan [103] & Yang [142] & Ours \\
\hline
baby & 1.8184 & 1.7387 & 2.3597 \\
bird & 1.9941 & 2.2013 & 2.5754 \\
butterfly & 2.3266 & 2.4080 & 2.5282 \\
head & 1.4845 & 1.4956 & 1.8091 \\
woman & 1.8733 & 1.9373 & 2.0316 \\
\hline
\end{tabular}
\caption{IFC comparisons with peer super-resolution methods on SET5 [9].}
\end{table}

\footnote{The results are generated using the executable file provided by the authors. We use the default parameters.}
“bird”, “butterfly”, “head”, and “woman”) under a scaling factor of 4 utilizing IFC. The high-resolution images for [142] are generated with the released code provided by the authors. Table 2.3 presents the corresponding comparison results. As observed, our method outperforms [103, 142] in all 5 images measured in IFC.

2.2 Internal Example-based Image Enhancement

Among the numerous situations of image enhancement, there are complicated circumstances where part of an input low-resolution image is undesired or unavailable, e.g., an overexposed spot, an unwanted shadow or scratch within a digital photo, or a satellite image partly covered by cloud. For these cases, in addition to the enhancement of image resolution, image completion techniques are also needed to reconstruct more visually plausible results. The underlying goal for both image super-resolution and completion is to predict the unknown pixel values within an image. They are numerically ill-posed problems and rely on additional assumptions or priors to finalize the output among all the possible solutions.

In this section, we extend the previous super-resolution framework which is based on gradient similarity and present a novel and straightforward algorithm for image enhancement which performs super-resolution and completion simultaneously. Provided an input low-resolution image and a mask representing the missing region(s), we perform enhancement in both gradient level and image level. The input low-resolution gradients in horizontal and vertical directions (represented as $x$ and $y$) are upscaled with completion embedded. The resulting high-resolution gradients, along with the completed low-resolution image, are fed into an optimization framework to reconstruct the final output image.
CHAPTER 2. INTERNAL EXAMPLE-BASED IMAGE ENHANCEMENT

47

Figure 2.10: Flowchart of the proposed internal example-based image enhancement framework. Given an input low-resolution image with a mask indicating the missing region(s), the low-resolution gradients are upsampled with completion embedded. The high-resolution gradients, along with the completed low-resolution image, are fed into an energy function to reconstruct the final high-resolution image.

2.2.1 Framework Details

Details of the proposed image enhancement framework is presented in this subsection which is based on internal exemplar similarity within the same scale and across different resolutions. Both gradient-level and image-level enhancement are employed to better preserve intensity changes and to ensure robust performance. Fig. 2.10 illustrates the schematic pipeline of our approach. Provided an input low-resolution image and a mask indicating the region(s) with missing pixels, the system consists of three components to accomplish both super-resolution and completion: gradient-level upscaling with completion, image-level completion, and the final high-resolution image reconstruction.

Gradient-Level Upscaling with Completion

Internal gradient similarity is utilized to simultaneously accomplish super-resolution and completion for input low-resolution gradients. It is based on the observation that since for small image patches in a natural image, self-similarities exist within the image itself and across different scales, we should expect this similar redundancy for gradient patches.

Given a grayscale input low-resolution image $L$, a mask $M$, and the scaling factor $s$, we
denote the gradients of $L$ in vertical and horizontal directions as $L_x$ and $L_y$. The enhanced high-resolution gradients are represented as $H_x$ and $H_y$. $L_x$ is decomposed into a set of overlapping patches with size $a \times a$. Patches with unknown pixels are upscaled first. Among the patches with missing regions, the upsampling priority for patch $P$ centered at pixel $q$ is determined as follows:

$$\text{Pri}(q) = C(q) \cdot D(q) = \frac{1}{a^2} \cdot \sum_{i \subseteq \{P(q) \cap \bar{M}\}} C(i) \cdot \sqrt{\frac{\nabla L_{xq} \cdot u_q}{N}}$$

where $P(q)$ represents the patch centered at pixel $q$, $\bar{M}$ indicates the unmasked region, $N$ is a normalization factor (255 for grayscale images), $u_q$ stands for a unit vector orthogonal to the front at pixel $q$. The initialization for $C(i)$ is set to $C(i) = 0$ if pixel value $i$ is unknown and $C(i) = 1$ otherwise. $L_{xq}$ represents the value at pixel $q$ in $L_x$. As indicated in Eq. 2.4, priority at a given pixel is measured as the product of two terms: the confidence term $C(\cdot)$ and the data term $D(\cdot)$. Both terms are normalized to range between 0 and 1.

Different from the priority computation in [18], we assign more credit to the data term during the calculation by modifying it to a squared form. The confidence term remains unchanged. In general, the confidence term measures the amount of reliable information surrounding a given pixel. The data term detects how strongly an isophote at that pixel collides and the contour at the same pixel.

After calculating the priority for each pixel along the boundary of the masked region, the patch with the highest priority at its center pixel is selected as the query patch $P$ to be upsampled. We then downsample $L_x$ by the scaling factor $s$ to obtain $LL_x$. A gradient patch pool $\mathcal{P}_x$ is formed with all the patches in $LL_x$ (size $a \times a$) whose pixel values are all known. To ensure a more expressive representation, all the patches in $\mathcal{P}_x$ as well as the query patch $P$ are normalized to have zero mean and unit variance.

Given a query patch $P$, its $k$ most similar patches are searched within the patch pool
The similarity between two patches is measured in mean square error with only the available pixels. After obtaining the $k$ similar patches in $LL_x$, their corresponding “parent” patches in $L_x$ are extracted and combined weightedly using Eq. 2.1. The combined patch is then readjusted according to the original mean and variance of $P$ and “pasted” to the corresponding position in $H_x$. After updating the confidence term and data term, the above process is repeated until all patches which have overlaps with the mask $M$ are upsampled. Then the rest patches in $L_x$ are upscaled in a similar manner as presented in Sec. 2.1. $H_y$ is computed in the same structure utilizing $L_y$.

**Image Level Completion**

To ensure a robust enhancement performance in constructing the final image, we also perform the image level completion on the input low-resolution image $L$ before the final reconstruction step.

Priorities for every pixel along the boundary of the masked region in $L$ are calculated according to Eq. 2.4. We then form a patch pool with all the patches (size $a \times a$) in the unmasked region of $L$. The $k$ most similar patches of patch $P$ whose center pixel has the highest priority are searched within the patch pool. Afterwards, the $k$ found patches are combined weightedly based on their similarity with the query patch. The unknown pixel values in patch $P$ are filled with the corresponding values in the combined patch.

After updating the confidence term and data term for the filled pixels, the above process is repeated until all the pixel values within the mask region are predicted. Finally, the completed image $L_C$ along with the high-resolution gradients $H_x$, $H_y$ are utilized for the final reconstruction of the target high-resolution image $H_C$. 
Final Image Reconstruction

After obtaining the completed low-resolution image $L_C$ and the high-resolution gradients $H_x$, $H_y$, the output high-resolution image $H_C$ is reconstructed through minimizing the same cost function shown in Eq. 2.2. Similarly, the first term ensures the consistency between the output high-resolution image and the completed input low-resolution image. The second term poses a constraint on the gradients of the target high-resolution image based on the gradients calculated after the ‘gradient-level upscaling with completion’ step. The cost function can be optimized through the gradient descent algorithm iteratively using Eq. 2.3.

As illustrated in Fig. 2.11, under different masks, the proposed framework well enhances the resolution of the input low-resolution image and predicts the missing pixel values in a way that is visually plausible.

2.2.2 Experimental Results

The proposed internal example-based image enhancement framework is evaluated with multiple natural images in different masks at a magnification factor of 2. Generally, patch size used in internal example-based super-resolution approaches is not large since patch recurrence occur among small patches. However, for patch-based image completion algorithms, we need a relatively large patch to retain local structure information. Therefore, in the proposed framework, to balance between these two tasks, we set the patch size $a$ to 7. In the calculation of pixel priority, normalization factor $N$ is 255 for grayscale images. During the nearest neighbor search, number $k$ of similar patches extracted is set to 10. The weighting factor $\lambda$ is 0.2 in the final image reconstruction step.

Fig. 2.12 presents the image enhancement result of image “snow”. As illustrated by the zoom-in region, the missing pixels are restored and upsampled with textures consistent with the overall structure. After image enhancement, sharp edges and fine textures are restored.
Figure 2.11: Enhancement results of image “fish” (×2) with two different masks. Left: input low-resolution images with missing regions (marked in white). Right: generated high-resolution results with clear contours and natural textures.
Figure 2.12: Enhancement result of image “snow” (×2). Left: input low quality image with missing region (marked in white). Right: output high quality images after enhancement.
Figure 2.13: Enhancement results of 4 images in BSDS dataset (×2) with different masks. Left: input low quality images with missing region(s) (marked in white). Right: output high quality images after enhancement.
Fig. 2.13 provides more results with 4 images in BSDS dataset of different masks. Enhancement for certain input low-quality images is challenging due to the existence of complicated textures, e.g., fur, trees, and complicated patterns. Zoom-in areas clearly indicate that our method correctly predicts the missing pixel values and reconstructs natural contours and fine details with minimal visual artifacts.

2.3 Discussions

In this Chapter, internal gradient patch similarity is utilized for single image super-resolution and completion tasks. No external statistics is used and therefore the image enhancement is accomplished solely based on exploring the internal statistics. The fidelity between the input and output images are ensured locally (i.e., through patch-based matching) and globally (i.e., with the explicit term in the cost function) in both image level and gradient level. The employed cost function nicely combines constraints from different domains into one uniform framework.

The limitations of the proposed super-resolution framework mainly lie in the increasing computational cost with the increment of the scaling factor and the degraded performance dealing with noisy input images. Another super-resolution framework is presented in Chapter 3 to address these issues.

Finally, regarding the quantitative evaluation of different super-resolution approaches – in [138], a variety of image quality metrics (i.e., PSNR, SSIM, VIF, MSSSIM, UIQI, WPSNR, IFC, and NQM) are evaluated by investigating the correlation with the visual perception by human experts. Although we still do not think currently there exists a widely-accepted systematic way to evaluate a super-resolution algorithm, among the 8 listed metrics, IFC [104] seems to be more proper to perform the comparison among different super-resolution approaches during the time the proposed algorithm is developed. On the other hand, in
Sec. 2.1.2 and 2.1.3, PSNR and SSIM are still used either to show the similarity between a generated output and the original image, or for comparison of the same framework at different patch pools or parameter settings.

In the later chapters, we occasionally use PSNR for the comparison with peer super-resolution methods (if PSNR is reported in the related publications) due to the fact that, although the misalignment between PSNR and visual evaluation is reported in several publications, it is still one of the most commonly used metrics in recent papers. To follow this mainstream and report our statistics, the PSNR values are listed to assist the visual evaluation. Moreover, since image super-resolution and completion target at generating visually plausible or pleasing results, qualitative evaluation is more essential for the comparison.
Chapter 3

Hybrid Example-based

Super-Resolution and its Application

Effectiveness of the internal statistics in single image super-resolution and completion has been demonstrated in Chapter 2. However, for image super-resolution, the difficulty in estimating missing high frequency details increases as the scaling factor gets larger due to the increment of low-resolution/high-resolution ambiguity. Moreover, it would be challenging to form expressive patch pools for input images with small sizes. On the other hand, external example-based learning breaks this limitation by introducing new information from a natural image dataset. As mentioned previously, the patch-to-patch similarity comparison can be time consuming when external statistics is involved. This problem can be solved if there is a way to avoid this patch-based comparison. However, the performance of external example-based super-resolution depends on the similarity between the training dataset and the testing images. Due to the diversification in natural images, the lack of relevance between certain testing images and a universal training dataset still exists. Keeping increasing the size of the training dataset provides a limited solution but still leaves the key problem untouched.
Figure 3.1: Super-resolution result of image “face” (×2). (a) High-resolution image generated by the proposed super-resolution method. (b) Ground-truth image. The proposed hybrid example-based super-resolution approach reconstructs natural and realistic edges and textures close to the original ground-truth image.

In this chapter\(^1\), we propose a novel hybrid example-based single image super-resolution method to address the limitations above. The framework incorporates learning image-level statistics from an external dataset followed by the gradient-level self-refinement with internal statistics. Afterwards, we demonstrate the effectiveness of the proposed framework by presenting a test case application of the super-resolution method enhancing passive microwave derived sea ice motions.

### 3.1 Hybrid Example-based Super-Resolution

The proposed hybrid example-based super-resolution scheme consists of three steps: a proxy high-resolution image is constructed through a set of pre-built regression models learned from external exemplars; the gradients of the proxy image are then fed into a pyramid self-awareness framework guided by the input low-resolution gradients; finally, the refined high-resolution gradients and the input image are integrated into a uniform cost function to

\(^1\)This chapter was previously published in [130, 134].
Figure 3.2: Flowchart of the hybrid example-based super-resolution method. Given a low-resolution image, a proxy high-resolution image is constructed through the regression models trained via an external dataset. The input feature space is modeled with GMM to ensure a targeted learning. With the proxy image, its gradients are refined using gradients of the input image. The refined high-resolution gradients are then integrated into the reconstruction framework to recover the final output image.

recover the final high-resolution image. Fig. 3.1 illustrates the comparison of the generated high-resolution image “face” and the ground-truth image. Our super-resolution result is very close to the original ground-truth image. Edge details including eye and face contours are natural and realistic. Hair textures are well reconstructed with minimal visual artifacts.

3.1.1 Framework Details

External example learning-based super-resolution usually relies on learning priors or models from a natural image dataset which in general leads to a stable super-resolution performance. Different from internal example-based approaches, learning externally can be performed offline and is less time consuming if the patch-to-patch comparison is not necessary during online super-resolution. However, natural images vary dramatically especially for edges
and textured regions. Given a natural image, certain patches occur rarely in the training
dataset and this results in a less effective super-resolution performance for those patterns.
On the other hand, internal patch redundancy has been validated to be effective both in
“expressiveness” and “predictive power” for image enhancement tasks [148]. In order to
combine the benefits of external and internal example-based learning, we propose a hybrid
learning-based super-resolution framework. Fig. 3.2 illustrates the schematic pipeline of
our approach. The system consists of three steps to upsample an image, i.e., proxy image
recovery from external statistics, gradient-level self-awareness from internal statistics, and
final image reconstruction.

Provided a low-resolution image, a proxy high-resolution image is first generated with a
group of pre-trained regression models. The regression models are trained on an external
natural image dataset. To ensure a targeted learning, the input feature space is modeled
with Gaussian Mixture Models (GMM) where an individual regression model is trained for
each Gaussian component. The generated proxy high-resolution image is robust with stable
super-resolution performance since the regression models are trained through a large number
of natural images in a divide-and-conquer manner. However, certain low-resolution patches
in the input image may appear rarely within the training dataset and thus lead to an in-
accurate high-resolution prediction, i.e., over-smoothed with missing high-frequency details.
Therefore, after obtaining the proxy image, a gradient-level coarse-to-fine self-refinement is
performed guided by gradients of the input image. Motivated by the reconstruction-based
super-resolution approaches, a gradient-level refinement is adopted to better preserve the
intensity changes. This process aims to replace the high-variance gradient patches in the
proxy image with more accurate representations to recover more visually plausible outputs.
Finally, the targeted high-resolution image is restored through minimizing a uniform cost
function with the refined gradients.
Proxy Image Recovery

Given an input image $L$, we first recover a proxy high-resolution image using a set of externally-trained regression models. A large set of low-resolution/high-resolution exemplar patch pairs with magnification factor $s$ is collected from a dataset consisted of more than 6,000 images. All images within the dataset are considered high-resolution images and the corresponding low-resolution images are generated with a blur and downsampling process.

To better preserve the structure information, for a low-resolution/high-resolution patch pair $\{P_l, P_h\}$, we normalize both patches by extracting the mean value of $P_l$. After normalization and vectorization, the input low-resolution and high-resolution features are represented as $X \in \mathbb{R}^{l \times M}$ and $Y \in \mathbb{R}^{r \times M}$ respectively where $l$ and $r$ denote the corresponding feature dimensions and $M$ indicates the number of samples.

To ensure a targeted learning, we first model the input low-resolution feature space where later multiple regression models are trained. The most straightforward model to describe the feature space is the normal distribution. However, a single normal distribution is insufficient to capture the complex nature of the features. We therefore employ GMM to represent the feature distribution. GMM is a generative model which has the capacity to model any given probability distribution function when the number of Gaussian components is large enough.

Given a GMM with $K$ components, the probability of a feature $x_i$ is

$$p(x_i|\theta_G) = \sum_{k=1}^{K} w_k \mathcal{N}(x_i; \mu_k, \sigma_k)$$

(3.1)

where $w_k$ is the prior mode probability which satisfies the constraint $\sum_{k=1}^{K} w_k = 1$, and $\mathcal{N}(x_i; \mu_k, \sigma_k)$ indicates the $k$-th normal distribution with mean $\mu_k$ and variance $\sigma_k$:

$$\mathcal{N}(x_i; \mu_k, \sigma_k) = \frac{\exp \left(-\frac{1}{2} (x_i - \mu_k)^T (\sigma_k)^{-1} (x_i - \mu_k) \right)}{(2\pi)^{l/2} |\sigma_k|^{l/2}}$$

(3.2)
where $x_i \in \mathbb{R}^l$, $\mu_k \in \mathbb{R}^l$, and $\sigma_k \in \mathbb{R}^{l\times l}$. By using the Expectation-Maximization (EM) algorithm to optimize the Maximum Likelihood (ML) from a large number of features, we can estimate the GMM parameters $\theta_G = \{w_k, \mu_k, \sigma_k, k = 1, \ldots, K\}$. We employ 200,000 randomly sampled features to learn the parameters $\theta_G$ in our experiment. Though Eq. 3.1 supports the full covariance matrix, a diagonal matrix in practice is sufficient to model most distributions. Moreover, GMM with diagonal matrices is more computationally efficient and stable compared to the one with full matrices.

GMM is based on a well-defined statistical model and is computationally tractable. We then assign each low-resolution feature $x_i \in X$ to corresponding Gaussian component with the highest probability. Suppose there are $M_k$ patches associated with the $k$-th Gaussian component and $X_k \in \mathbb{R}^{l \times M_k}$, $Y_k \in \mathbb{R}^{r \times M_k}$ represent the corresponding low-resolution/high-resolution features, a linear regression model is then trained with the regression coefficient $A_k$ learnt through:

$$A_k^* = \arg\min_{A_k} \{|Y_k - A_k \hat{X}_k|^2\}$$

(3.3)

where $\hat{X}_k = [X_k^T \mathbf{1}]^T$. During the testing phase, given a low-resolution image, we first extract the features by performing normalization and vectorization for every low-resolution patch. Then each feature is assigned to a Gaussian component according to the posterior where the corresponding regression model is applied to obtain the high-resolution patch. We use simple average to blend overlapping pixels in generating the proxy high-resolution image.

**Gradient-Level Self-Awareness**

Internal patch redundancy has been demonstrated powerful for image restoration tasks [148] and serves as the theoretical foundation for internal example-based super-resolution methods. With good performance for super-resolution under relatively small magnification factors, the
Figure 3.3: Variance distributions of 20,000 patches with size $7 \times 7$ extracted from BSDS dataset. Given a patch, the larger the variance is, the less frequent it tends to appear within the dataset.

The limitations of internal example-based super-resolution approaches lie in the heavy computational costs to execute online exhaustive pair-wise patch comparisons and the degraded performance with the increment of the scaling factor. Although the aforementioned internal gradient similarity-based super-resolution method is capable to handle large scaling factors, the computational cost increases dramatically with the increment of the scaling factors.

In this step, we aim to absorb the advantages of self-similarity to refine the proxy image generated previously without going through the exhaustive patch matching.

The self refinement process aims at recovering the missing high-frequency details for patches which are not frequently seen in the external training dataset. We first verify that patches with higher variances tend to appear less frequently within a natural image dataset. The experiment is performed by randomly selecting 20,000 patches of size $7 \times 7$ within the BSDS dataset [79]. As observed from Fig. 3.3, the number of patch instances decreases quickly as the variance increases.
Figure 3.4: (a) Illustration of the gradient-level coarse-to-fine self-awareness procedure. Gradients of the proxy image $H_{(x,y)}^p$ are downsampled to $M_{(x,y)}^p$ which are refined with the gradients $L_{(x,y)}$ of the input image. Gradients with darker frames represent the corresponding refined results of the ones with lighter frames. Afterwards, $H_{(x,y)}^p$ is refined with $M_{(x,y)}$. Please refer to text for details. (b) Difference map between $H_{(x)}$ and $H_{(x)}^p$. (c) Difference map between $H_{(y)}$ and $H_{(y)}^p$. 
To refine $H^p$ with $L$, patches from $H^p$ of size $a \times a$ with variance larger than a preset threshold $\theta_r$ are extracted first. We utilize self-similarity to recover the missing high-frequency details of those high variance patches which may not be frequently seen in the training dataset. For each high variance patch, its $k$ most similar patches with the same size are searched and extracted within $L$ and the similarity of two patches is measured in their mean square error. Afterwards, the original patch is replaced with the weighted sum of the found $k$ patches in a softmax way.

In the proposed scheme, a gradient-level self-refinement is adopted to better preserve edge and texture information. Moreover, it is validated in [38] that average patch recurrence across scale decays as the resolution difference increases. Therefore, if the magnification factor $s$ is larger than $s_0$ ($s_0 = 3$ in our experimental setting), the proposed self-refinement is executed in a coarse-to-fine scheme.

Fig. 3.4 illustrates the self-awareness process. After obtaining the proxy image $H^p$, its gradients in horizontal and vertical (denoted as $x$ and $y$) directions are computed and refined with the corresponding gradients of the input image $L$. In later context, for ease of interpretation, we denote the gradients of an image $I$ in $x$ and $y$ directions as $I_{\{x,y\}}$. The refinement is performed separately for each gradient direction. Taking the horizontal direction as an example, if $s$ is larger than $s_0$, we first downsample $H^p_x$ by factor $\sqrt{s}$ to obtain $M^p_x$. After that, high variance patches in $M^p_x$ are refined using $L_x$ to obtain finer-version gradient $M_x$. Then the final high-resolution gradient $H_x$ is computed by utilizing $M_x$ to refine $H^p_x$. In the above process, if $\sqrt{s}$ is still larger than $s_0$, we further decompose $\sqrt{s}$ in a similar manner.

Gradient patches are mostly flat with small variances. Therefore, only a small portion of the patches are refined to recover the missing high frequency details. To ensure an effective refinement, all the patches are normalized to have zero means and unit standard variances before searching. The combined patch is then readjusted according to the original mean
CHAPTER 3. HYBRID EXAMPLE-BASED SUPER-RESOLUTION

Figure 3.5: Comparisons measured in average PSNR (dB) between the proxy images (marked in blue) and the corresponding final output (marked in red) recovered from self-refinement and reconstruction in datasets BSDS [79], SET5 [9], and SET14 [145] (scaling factor $\times 4$). There is an obvious boost in PSNR after performing self-awareness and reconstruction for all three datasets.

and variance of the input patch. After the self-awareness step, over-smoothed patches in the proxy high-resolution image are refined.

Final Image Reconstruction

The final step is to reconstruct the output image from the self-refined high-resolution gradients through the cost function which is the same as Eq. 2.2 (it is listed here again for the ease of reading):

$$H^* = \arg\min_H \{|\nabla H - \nabla H_r|^2 + \lambda|(H \ast G) \downarrow_s - L|^2\},$$

(3.4)

where $\nabla H_r$ represents the refined $H_{x,y}$ after the pyramid gradient-level self-awareness step. $G$ stands for a Gaussian kernel with standard variance $\sigma$ varies for different scaling factors $s$: $\sigma = \{0.8, 1.2, 1.6\}$ for $s = \{2, 3, 4\}$. $\lambda$ is the weighting factor. The cost function can be optimized through the gradient descent algorithm (details in Eq. 2.3).
We demonstrate the effectiveness of the proposed gradient-level self-refinement and the feasibility of reconstructing images based on gradients experimentally on datasets BSDS [79] (200 images), SET5 [9] (5 images), and SET14 [145] (14 images). All the images are downsampled by a factor of 4. Fig. 3.5 presents the average PSNR (dB) comparison of the proxy images and the final outputs after self-awareness and reconstruction within each dataset. After refining the ambiguous patches, all the images experience an obvious boost in the super-resolution performance measured in PSNR.

### 3.1.2 Experimental Results

The proposed hybrid example-based super-resolution method is evaluated with multiple images on datasets SET5 [9], SET14 [145], and BSDS [79]. We compare our results with state-of-the-art single image super-resolution algorithms both quantitatively and qualitatively.

**Parameter Selection:** Same as many existing super-resolution methods, for color images, the proposed algorithm is applied on the luminance channel in the YUV color space while the other two color channels are upsampled with bicubic interpolation.

The training dataset used for regression model learning is the same as in [139] with 6,152 natural images. We extract all patches with size $7 \times 7$ from the generated low-resolution images. Corners of each patch are removed and thus the low-resolution feature dimension is 45. Only the central $3s \times 3s$ pixels in the corresponding high-resolution patch are used to formulate the high-resolution feature where $s$ indicates the magnification factor. We randomly select 200,000 low-resolution/high-resolution features to train the GMM with 512 components. To better model the feature space, we filter out the smooth patches before selection. With the trained GMM, each feature is assigned to a Gaussian component with the highest probability. A linear regression model is learned for each Gaussian component.
Table 3.1: Comparison of the proposed approach with recent state-of-the-art methods in SET5 [9], SET14 [145], and BSDS [79] in terms of average PSNR (dB). Our results outperform other methods in all three datasets.

|---------|---------|------------|------------|------------|-----------|------------|------|

using maximum 1,000 low-resolution/high-resolution features within this component.

In the pyramid gradient-level self-awareness step, the maximum magnification factor \( s_0 \) between each level is 3. If the scaling factor \( s \) is larger than 3, we adopt a coarse-to-fine scheme with a factor of \( \sqrt{s} \) per-step. Patch size \( a \) in the self-refinement is 7 and the pre-set threshold \( \theta \), used to differentiate the smooth patches with the high-variance ones is 5. The number \( k \) of similar patches captured during the searching is 5. We set \( \lambda \) in Eq. 3.4 to be 4/7.

Quantitative Analysis: The proposed approach is evaluated on a variety of natural images and we compare the generated results with recent state-of-the-art methods [103, 120, 139, 141, 145] quantitatively measured in PSNR. We use the source code [120, 139, 145] and executable file [103] provided by the authors or a third-party implementation [141] to generate the corresponding high-resolution images based on the same low-resolution input images. To be more specific, given a ground-truth image, the low-resolution image is obtained by performing the bicubic downsampling. These low-resolution images are saved in PNG format and serve as the uniform input for all super-resolution approaches. We then follow the code or executable file provided by the authors to perform the image upsampling. Super-resolution methods [120, 141, 145] generate results with borders shaved. To perform a fair comparison, we crop the borders for all the high-resolution results generated by different super-resolution approaches utilizing the same scheme before the PSNR calculation over the luma channel.

The evaluation is performed on three datasets, i.e., SET5 [9], SET14 [145], and BSDS
Figure 3.6: Super-resolution of image “Lenna” (×4). Zoom-ins clearly indicates that the proposed hybrid example-based super-resolution framework reconstructs the hat contours with minimal artifacts while other methods suffer from blurring, jaggies, or aliasing artifacts.
CHAPTER 3. HYBRID EXAMPLE-BASED SUPER-RESOLUTION

Figure 3.7: Super-resolution of image “snow” (×4). It is a challenging task to reconstruct the rattan textures. Results generated by bicubic interpolation and [142] over-smooth the textures. Deformed patterns exist in [103] (squared textures) and [139] (discontinuities). [120] fails to recover several edges as shown in the circled zoom-in and [24] oversharps the edges. Our result best reconstructs the details.
[79] at magnification factor 4. As illustrated in Table 3.1, for different input images, the proposed approach outperforms the other methods measured in average PSNR (dB) over all three datasets.

**Qualitative Analysis:** Fig. 3.6 presents a set of super-resolution results on image “Lenna” with an upscaling factor of 4. Reconstructing the hat contours without obvious artifacts is difficult for most of the peer super-resolution approaches. Our method successfully generates clear contours consistent with the ground-truth image.

Fig. 3.7 provides another set of results with a scaling factor of 4 on image “snow”. It is challenging to reconstruct the rattan textures as illustrated by the zoom-in regions. Results generated by bicubic interpolation and [142] are over-smoothed. Deformed patterns exist within [103] (irregular squared patterns) and [139] (discontinuities). [120] fails to recover several edges and [24] over-sharps the edges. Image generated by hybrid example-based super-resolution recovers more natural patterns.

### 3.2 Super-Resolved Fine Scale Sea Ice Motion Tracking

Sea ice is a vital component in the Earth’s climate as well as posing potential hazards to shipping and other maritime activities in the Polar regions. It is particularly critical to monitor and track the motions of sea ice in near real-time (i.e., within several hours of data acquisition) for safe naval operations in the Arctic Ocean, as well as to further validate or improve models of the polar ice pack, coupled with predictors like ocean temperatures, sea level pressure, and geostrophic winds, for ice hazard forecasts at a finer scale [14, 57, 82, 107]. Several international operational ice centers provide routine tactical and strategic ice analyses in support of navigation and other activities in the Arctic, including the U.S. National Ice
Center in Suitland, Maryland, USA [123].

The majority of operational sea ice monitoring techniques relies on satellite-borne optical and synthetic aperture radar (SAR) sensors, augmented by scatterometer and passive microwave imagery [8, 93]. Feasibility and accuracy in ice motion tracking hinge on the spatial and temporal resolutions of the input data. High spatial resolution (i.e., 100 − 1,000 m) is possible with visible and SAR imagery, but at the cost of limited temporal sampling due to clouds (for visible imagery) or limited coverage (for SAR—narrow swaths, longer orbital repeat visits). On the other hand, passive microwave data can provide near-complete daily coverage over the entire Arctic, but at low spatial resolutions (i.e., 12.5 − 25 km) [63, 81, 82].

Compared with other means of measuring ice drift (e.g., buoys), satellite sensors provide a more complete and routine coverage of Polar regions [37]. In particular, with the capabilities to penetrate cloud cover and observe the surface all day, satellite microwave sensors, including passive microwave sensors, are often considered as the best option to estimate sea ice drifts. The Defense Meteorological Satellite Program (DMSP) Special Sensor Microwave Radiometer and its predecessors, the Nimbus-7 Scanning Multichannel Microwave Radiometer (SMMR) and a series of DMSP Special Sensor Microwave/Imager (SSM/I) and Special Sensor Microwave Imager and Sounder (SSMIS) instruments have been operated for over 30 years, providing a long time-series of sea ice motion data [121].

Beginning in 2002, the National Aeronautics and Space Administration (NASA) Earth Observing System (EOS) Advanced Microwave Scanning Radiometer (AMSR-E) on the Aqua platform and the Japan Aerospace Exploration Agency (JAXA) AMSR2 sensor on the Global Change Observation Mission-Water (GCOM-W) platform have provided higher resolution passive microwave imagery. The higher spatial resolution (i.e., 5 − 12.5 km) yields more accurate estimates of sea ice motions [62, 80], though fine scale motions (such as fractures and small lead openings) are still not detectable. These fine scale motions are important for effectively tracking energy fluxes, ice growth, and ocean freshwater fluxes. Moreover, they
are critically important for navigational guidance in ice-infested waters. Providing motions at the passive microwave spatial and temporal coverage but at enhanced resolution will be a significant benefit.

In this section, we present a test case application of image super-resolution method enhancing passive microwave derived sea ice motions. Super-resolution techniques have been utilized in the remote sensing field for various applications [41, 70, 83]. We aim to accurately track sea ice motion at fine scales by first constructing high-resolution images with the aforementioned hybrid example-based super-resolution algorithm. Afterwards, a benchmark tracking algorithm based on maximum cross-correlation (MCC) [28] is applied to estimate sea ice drift vectors and track the sea ice movements. We adopt the benchmark MCC algorithm since in principle, it can work on any type of imagery and has been successfully applied to sea ice with visible/infrared, scatterometer, SAR, and passive microwave data. To demonstrate the potential of the super-resolution algorithm in further relevant applications, tracking of specific individual sea ice objects is additionally demonstrated using a state-of-the-art object tracking algorithm [23]. Overall, this section shows that by using the super-resolved images, the accuracy of sea ice drift estimation is significantly improved compared to using the original images.

Sea ice motion through remote sensing data has been extensively studied since the appearance and wide availability of satellite imagery. As a side-note, although the term “motion” may strictly refer to continuous monitoring (e.g., through video), while imagery practically allows to detect displacements, or drifts, due to its common use in the literature, the term is used in this section interchangeably with the term “drift”. Sea ice drift is usually estimated with satellite imagery via a pattern matching method (e.g., [28, 75]). Sea ice drift vectors have been derived from a variety of satellite imagery utilizing a MCC criterion [27, 80, 82]. These methods have been used successfully for sea ice motion for a variety of imagery, including visible [28], SAR [63, 110], scatterometer [42], and passive microwave [82].
Figure 3.8: Daily composite AMSR2 image at 36.5 GHz from January 1, 2013, projected on a polar stereographic grid. The latitude-longitude coordinates of the image corners are: upper-left (30.98°, 168.35°), upper-right (31.37°, 102.34°), lower-left (33.92°, 279.26°), lower-right (34.35°, 350.03°) [87].
A key factor in the pattern matching methods is that a characteristic and stable pattern, or a highly similar pattern, needs to be detected in both images of a pair to potentially retrieve a motion estimate. Visible and infrared sensors are limited by clouds, which are often prevalent in the Arctic regions. This substantially limits the number of motion vectors that can be retrieved. Likewise, other high-resolution sensors, such as SAR, have narrow swath widths and limited repeat coverage. Though having a much lower spatial resolution, passive microwave imagery has been especially useful for monitoring sea ice motion because it is independent of solar radiation, has complete daily coverage of the Arctic regions, and atmospheric interference is insignificant in most cases.

MCC-based approaches have been widely applied recently in sea ice drift estimation with several variations. Thomas et al. [118] proposed a method for sea ice motion characterization at a 400-m resolution vector field using European Remote Sensing Satellite-1 (ERS-1) SAR imagery. Motion fields of sea ice are obtained utilizing Phase Correlation (PC) pre-selection and MCC in a multi-resolution processing system. In [119], Thomas et al. developed a sea ice motion tracking system at the geospatial mesoscale (i.e., 1 – 100 km²) and proposed an adaptation of the algorithm that estimates drifts at close proximity to discontinuous regions using image inpainting. Building on this approach, Hollands and Dierking [49] implemented a PC- and MCC-based pattern matching algorithm to identify corresponding sea ice structure in a sequence of SAR images for the observation of high-resolution sea ice motions in the Weddell Sea at spatial resolutions varying from a few hundred meters to a few kilometers. Adapting the pattern matching approach in [118] by adding a Fourier-Mellin transform to capture rotational motion, Berg and Eriksson [6] recently proposed a hybrid pattern matching and feature tracking approach, the latter component requiring an image segmentation pre-processing step. Komarov and Barber [59] introduced an approach for automated selection of control points to which PC was applied to estimate candidate translational and rotational drifts, followed by MCC for the final decision. In [65], Lavergne et al. had earlier introduced
a MCC-based sea ice motion tracking framework with a continuous optimization step for computing the motion vectors which are able to effectively reduce the quantization noise generated by MCC. This approach is proved capable of retrieving spatially smooth 48-h sea ice motion vector fields in the Arctic.

3.2.1 Data

We employ AMSR2 data acquired from the JAXA Earth Observation Research Center\(^2\). Level 2 swath data of horizontal polarization 36.5 GHz brightness temperatures are gridded on a 12.5 km polar stereographic grid, tangent to the Earth’s surface at 70 degrees northern latitude [87], using a simple drop-in-the-bucket method. All swaths from each day are averaged to create daily-average brightness temperature fields. The drift estimation algorithm is then applied to the gridded brightness temperatures. Seven daily such images from January 1-7, 2013 are employed, with the first one shown as an example in Fig. 3.8.

To further demonstrate the applicability of the proposed super-resolution framework in tracking sea ice motion, a number of images derived from the Visible Infrared Imaging Radiometer Suite (VIIRS) on board the Suomi National Polar-orbiting Partnership (NPP) satellite\(^3\) are used. The images are retrieved as a sequence of daily non-geolocated frames forming a video demonstrating the formation and motion of sea ice leads (cracks) around the Beaufort Sea along the northern coasts of Alaska and Canada. They include a total of 40 frames spanning from February 17 to March 18, 2013.

3.2.2 Framework Details

To enhance the quality of the satellite imagery and to further increase the sea ice motion tracking accuracy, we propose a sequential super-resolved sea ice motion tracking frame-
work. The high-resolution satellite imagery is firstly constructed utilizing the aforementioned hybrid example-based super-resolution algorithm within which both external and internal statistics contribute to recover quality edges and fine details. Afterwards, a benchmark drift estimation algorithm is applied to show the superiority of performing motion tracking with the super-resolved finer scale imagery. In an additional experimental setting, a state-of-the-art object tracking algorithm is applied on super-resolved images to track individual sea ice objects (i.e., floes). Details of the super-resolution method can be found in Sec. 3.1. The rest of the framework is presented in the following subsections.

Sea Ice Drift Estimation

The MCC method essentially matches patterns (e.g., grid cells) in two coincident images separated by a time interval through the use of a sliding window within a given neighborhood of the pattern’s location in the first image. The new location of the pattern in the second image is determined by searching for the location of the sliding window where the cross-correlation with the pattern in the first image is maximized. The motion estimate can be calculated in a straightforward manner by dividing the displacement distance by the time separation between the two images. In the proposed framework, the MCC-based method proposed in [28] is applied for drift estimation. Besides the Advanced Very High Resolution Radiometer (AVHRR) data used in [28], this method has been used with AMSR-E data in [80] and operationally applied to passive microwave imagery as a component in a sea ice motion product distributed by the National Snow and Ice Data Center (NSIDC) [121].

The spatial resolution is a limiting factor in drift estimation accuracy since in theory the displacement can only be determined in discrete increments corresponding to the grid resolution. In the method employed, an oversampling procedure is conducted to calculate motion at subpixel resolutions by moving the sliding window in increments of 1/4 of a grid cell instead of complete grid cells, in each direction [81, 82]. This is performed by implicitly
applying linear interpolation that sets the value of the sub-grid cells equal to the weighted average of the original ones, e.g., for a 1/4 sub-grid cell, it weighs the original grid cell 75% and the next grid cell 25%. With this oversampling procedure, sea ice drifts are expressed as displacements of products of the 1/4 of the original grid resolution, so the motion vector field resolution, i.e., the minimum expressed displacement is 3,125 m, or 3.62 cm s\(^{-1}\) for sequential day images. Thus, a drift larger than half of one sub-grid cell, i.e., 1/8 of an image pixel or 1,562.5 m, will be able to dominate the sub-grid cell value and be detected as drift by the algorithm—though, expressed as the minimum possible 3,125 m drift. Besides oversampling, other subpixel motion estimation approaches have been applied in the literature as part of MCC-based motion estimation, such as curve fitting in the correlation value domain \[63\] or interpolation in the image data \[65\]. They could be used instead of oversampling in the MCC-based motion estimation on the images generated after applying the super-resolution algorithm, i.e., still in tandem with the super-resolution approach, and comparison of them could be a topic for future research. In the proposed framework, staying consistent with the algorithm as applied in \[28\], \[80\], and \[121\], we use a 4× oversampling as described above.

Post-processing quality control of the estimated motion vectors is then performed via two filtering approaches. First, a minimum correlation threshold is applied to remove marginal pattern matches that are more likely to be in error. Thresholds of 0.1 and 0.4 were selected in \[28\] for AVHRR data, 0.6 in \[37\] for scatterometer data, whereas 0.7 in \[80\] for 36.5 and 89 GHz AMSR-E data; in our case, a threshold of 0.6 is selected after experimentation. Second, a spatial coherence filter is used to remove outlying displacements by comparing each motion vector with neighboring vectors. If one vector is an outlier (i.e., the number of neighboring vectors whose displacements are within 2 pixels of this vector is less than two), it is deemed to be erroneous and should be removed. This is effective because generally large-scale motion is correlated at distances up to several hundred kilometers.

Motion errors are dependent on several factors, such as the geolocation accuracy of the
input imagery, the validity of the assumption that the surface properties do not change between images, atmospheric interference, and the spatial resolution of the imagery. Detecting motion during summer is challenging because of surface melt and more atmospheric emission by the moister atmosphere; thus, here we focus on winter scenes. As in [28], a land mask is employed to identify and exclude all land pixels from the motion vector computations.

While each of the above contributes to the total error, the most significant limitation in the accuracy of ice motions is the spatial resolution of the source imagery. With the AMSR2 36.5 GHz data mapped to a 12.5 km resolution grid, the motion can only be detected if it moves at least one half grid cell during the chosen time interval (one half sub-grid cell when oversampling is performed, as detailed above). Higher frequency imagery, at 89 GHz, can also be used and has roughly double of the spatial resolution, gridded at 6.25 km. However, the 89 GHz imagery is more susceptible to atmospheric interference and for AMSR-E, the 36.5 GHz channels yield motion errors of a similar magnitude as the 89 GHz channels [80]. There is also uncertainty due to the use of daily average passive microwave images, where all swaths over 24 hours are averaged into a daily composite. This results in an ambiguous time interval because a 24-hour separation is assumed for all grid cells in the images. Thus, the retrieved pattern displacements are not associated with a distinct instant in time and result in a temporal smearing of the ice signal and distortion of surface patterns that inhibits correlation comparisons between days. However, daily composite microwave images have been preferred in several research studies [27, 80, 82] and operational products [121], since they reduce missing brightness temperature values and allow drifts to be calculated at the whole grid.

Fortunately, many errors are independent, including estimates from the same location at different times. This means that while individual vector estimates may have large errors, on average the errors are much smaller than the theoretical error and, importantly, the estimates are largely unbiased. For AMSR-E, the root mean square (RMS) error of daily motion speed
has been found to be on the order of 6 cm s\(^{-1}\) with directional RMS values on the order of 15 – 20 degrees [62, 80, 110].

A 6 cm s\(^{-1}\) is a reasonable uncertainty for looking at large scale sea ice circulation, particularly when tracking ice over several days or weeks. Under this circumstance, it is not possible to detect fine scale motions such as lead formation (openings in the ice) and ridging (convergent motion), which occur at scales of 1 km or less over subdaily intervals (corresponding to a speed of less than 1 cm s\(^{-1}\) over a day). These fine scale motions are important for local and regional processes (e.g., energy fluxes between the ocean and atmosphere) and for navigational guidance.

**Sea Ice Object Tracking**

In addition to the estimation of sea ice drifts, we further evaluate the potential of the images generated by the proposed super-resolution approach to track specific sea ice moving objects (i.e., floes). Whereas the outcome of the MCC-based drift estimation method is a set of motion vectors representing the displacement or velocity of sea ice on the image grid, the outcome of the tracking method is a bounding box indicating the position of a specified object on each frame of a series of images.

Having as input the sequence of the Suomi NPP images, as described in Section 3.2.1, we select a sea ice object in the first frame, i.e., the first image of the sequence, by manually defining the rectangular bounding box enclosing the object. Then we apply the context tracker algorithm [23] to identify the position of the sea ice object on the following images of the sequence. The algorithm has proven effective in computer vision tracking applications in unconstrained environments. One main reason for its selection here is its performance in tracking an object with the presence of similar neighboring ones that might cause confusion, e.g., similarly looking sea ice objects, and small changes in appearance. It works by defining a set of so-called distracters and supporters. Distracters are regions with appearance similar
to the tracking object that co-occur with it in several frames and might be potential sources of confusion. The algorithm tracks the distracters in addition to the selected object to prevent such confusion and false detection, even if the object is occluded in some frames. The supporters are the extracted key points around the targeted object which move together with it and help the tracking. Both of them are automatically explored using a sequential randomized forest, an online template-based appearance model, and local features.

### 3.2.3 Experimental Results

The proposed sequential super-resolved fine scale sea ice motion estimation system is evaluated with a sequence of passive microwave images. The dataset used for motion estimation is publicly available to allow reproduction or comparison of the results\(^4\). We also test the hybrid example-based super-resolution method on tracking three selected sea ice objects in a sequence of images from the Suomi NPP satellite. In both scenarios, there is an obvious boost in the tracking performance upon the super-resolved images.

#### Drift Estimation

To demonstrate the effectiveness of the super-resolved fine scale sea ice motion tracking, the proposed framework is evaluated over the sequence of the AMSR2 passive microwave images from January 1-7, 2013. Provided the ground-truth (original AMSR2) imagery, we first generate low-resolution inputs of 50 km pixel spacing, by downsampling them under a scaling factor of 4. After that, the hybrid example-based super-resolution algorithm is applied on the low-resolution images to generate the high-resolution instances with the same magnification factor 4. It is noteworthy that in a real application, the super-resolution process would be applied on the original images, creating new images of 3.125 km spatial resolution. The reason of applying the super-resolution process on the 50 km downsampled data and not the

\(^4\) [http://media-lab.ccny.cuny.edu/wordpress/Code/sea_ice_flow_dataset.zip](http://media-lab.ccny.cuny.edu/wordpress/Code/sea_ice_flow_dataset.zip)
CHAPTER 3. HYBRID EXAMPLE-BASED SUPER-RESOLUTION

Figure 3.9: Super-resolution of the passive microwave imagery (×4). From left to right, the columns represent zoom-ins of the ground-truth image, the low-resolution image, and the generated high-resolution image. It is clearly indicated that the hybrid example-based super-resolution approach reconstructs clearer ice contours consistent with the ground-truth instances.

The 12.5 km original ones, is the lack of images of spatial resolution higher than 12.5 km that we could use as ground-truth to evaluate the super-resolved images against. Since only 12.5 km resolution images are available, we use these as ground-truth and create images of the same resolution, to demonstrate how accurately the super-resolution process can increase the resolution of the input images (i.e., the 50 km ones) by 4 times and approximate the ideal high-resolution images (i.e., the 12.5 km original AMSR2 images). For ease of interpretation and direct comparison with the super-resolved imagery, we upscale the 50 km input images to the same size as the ground-truth images using nearest-neighbor interpolation and use the interpolated results as the representation of the low-resolution inputs to perform motion estimation.

Fig. 3.9 presents the zoom-in comparisons before and after the hybrid example-based super-resolution algorithm. Compared with the ground-truth instances, fine details and
Figure 3.10: Estimated motion vectors from the first pair (January 1-2, 2013) of images, for each of the original AMSR2 (Orig.), low-resolution nearest-neighbor (NN), and super-resolved (SR) image sets.
Following the generation of the nearest-neighbor interpolated (NN) and super-resolved (SR) images, the MCC-based drift estimation algorithm is applied on all pairs of consecutive days from each of the original AMSR2, NN, and SR image sets. Fig. 3.10 draws an example of the motion vector fields (i.e., velocities) resulting from the three image sets for the image pairs of January 1 and 2. The results cover the whole Arctic area and they are restricted to areas with sea ice. Areas of land and areas of sea and ocean not covered by ice are masked out. Zero-magnitude vectors are not drawn in the figure. It is readily seen that zero-magnitude vectors are much more common in the NN vector field than the SR field, i.e., the algorithm cannot detect sea ice drift in the NN image pair in the extent it does in the SR pair. It is also noted that the magnitude of the vectors estimated with the SR pair is closely related with the vectors from the original image pair. On the contrary, several vectors from the NN pair appear an order of magnitude larger than the original image pair vectors, even exceeding values of $70 \text{ cm s}^{-1}$ [121] and $100 \text{ cm s}^{-1}$ [69] considered as maximum realistic velocities. In particular, whereas the maximum velocities estimated for the original and SR pairs are $36.9 \text{ cm s}^{-1}$ and $50.1 \text{ cm s}^{-1}$, respectively, the maximum one for the NN pair is $157.1 \text{ cm s}^{-1}$. The results are similar for the image pairs from the rest studied days.

Fig. 3.11 offers a close-up look of Fig. 3.10 for the main Arctic region around the North Pole. The motion vector field from each image set appears in a separate figure. In addition to the observations discussed above, Fig. 3.11 highlights the similarities in the distribution of the non-zero-magnitude vectors between the original (Fig. 3.11(a)) and the SR (Fig. 3.11(c)) images. In a large area around the North Pole (black dot near the center of the images), the SR images are able to reveal drifts in a much closer detail than the NN images, which are incapable of depicting small sea ice displacements. On the contrary, the NN images result in the erroneous detection of large-extent drifts in a region where the two other image sets detect small or no drifts (western part of the region, Fig. 3.11(b)). This suggests the
existence of intense image artifacts in the NN images, in contrast with the smoother results by the SR images. As a note, since all land pixels are excluded from the motion vectors calculations in all images, any artifacts in the NN and SR images in the land pixels, and to a large degree in the coastlines, are expected to have only small effect in the estimated motion vectors.

To quantitatively evaluate the benefits from the super-resolved images compared with the low-resolution ones, Fig. 3.12 draws the scatterplots of the $x$-axis (vertical) and $y$-axis (horizontal) drifts from the two image sets, compared with the respective drifts from the original AMSR2 images considered as the ground-truth, for the image pairs on January 1 and 2. Several outliers can be noticed for the NN drift vectors in both axes (Fig. 3.12(a) and 3.12(b)), whereas the distribution of data in the SR is more compact. In addition, the least squares linear regression fit line is calculated for each image pair and axis drift and drawn as a solid line; the 1-by-1 ideal correlation between the NN or SR results and the original data is drawn as a dashed line. As observed, the fitting lines for the drifts in the SR images (Fig. 3.12(c) and 3.12(d)) are closer to the ideal-fit line than the one in the NN images (Fig. 3.12(a) and 3.12(b)), revealing that the correlation between SR and original
Figure 3.12: Scatterplots of NN and SR drift estimates (in km) in the two axes compared with the vectors from the original images considered as the ground-truth, for the image pairs on January 1 and 2. (a) NN drift in the $x$-axis (vertical); (b) NN drift in the $y$-axis (horizontal); (c) SR drift in the $x$-axis; (d) SR drift in the $y$-axis. In addition to the drift data, the 1-by-1 ideal match line is drawn as a dashed line, as well as the least squares linear regression fit line as a solid one.
image vectors is higher compared with the correlation between the NN and original image vectors.

Table 3.2 provides a thorough quantitative evaluation of the NN and SR motion vectors compared with the original image data considered as the ground-truth. As observed, the relative squared error (RSE), root mean squared error (RMSE), and mean absolute error (MAE) are consistently smaller for the SR vectors than the NN ones for both the vertical and horizontal drifts among all image pairs. On the contrary, the Pearson correlation coefficient (P) is significantly higher for SR vectors compared with the NN vectors. This shows that there is a strong positive correlation between the SR and original vectors in several cases, whereas on the contrary, almost no, or even slightly inverse (for Jan 2-3 pair and y-axis), correlation appears for NN vectors.

Fig. 3.13 plots the error distributions (in km) of the NN and SR motion vectors in the two axes, compared with the vectors from the original images, for the image pair Jan 1-2. Apart from a slight bias on the positive direction in the x axis, no significant biases are observed in the two vectors sets. However, a number of extreme error values (outliers) can be observed in both axes of the NN vectors. This seems to partially affect the standard deviation of the distribution which appears larger than the SR respective one for both axes in the schematic representation of the probability density function histograms appearing below and on the right of the main plot areas. In fact, the standard deviations of the errors for the SR vectors are consistently lower for all image pairs than the NN vectors, as seen in Table 3.3. This reveals the lack of outlier vectors in the SR images and the close relevance with the vectors from the original data. The mean values of the errors are closer to zero for the SR vectors than the NN vectors, showing a smaller bias towards negative or positive drifts. Regarding differences between the x and y directions, no strong biases in one versus the other direction are observed in either the NN or the SR data.
### Table 3.2: Quantitative evaluation of the NN and SR drift vectors compared with the vectors from the original AMSR2 images, for all image pairs individually and aggregated. "δx" and "δy" indicate the drifts on the vertical and horizontal axes, respectively. "Samples" stands for the total number of drift vectors compared, "RSE" for the relative squared error, "RMSE" the root mean squared error in km, "MAE" the mean absolute error in km, and "P" the Pearson correlation coefficient.

| Dates | Samples | δx | RSE  | RMSE | MAE  | P   | δy | RSE  | RMSE | MAE  | P   |
|-------|---------|----|------|------|------|-----|----|------|------|------|-----|-----|
| NN    |         |    |      |      |      |     |    |      |      |      |     |     |
| Jan 1–2 | 1165 | 6.16 | 12.22 | 3.82 | 0.04 | 6.94 | 12.28 | 4.08 | 0.02 |
| Jan 2–3 | 1163 | 5.03 | 13.14 | 4.66 | 0.09 | 3.34 | 12.73 | 4.79 | -0.01 |
| Jan 3–4 | 1120 | 3.34 | 11.26 | 4.10 | 0.12 | 2.80 | 12.54 | 4.87 | 0.04 |
| Jan 4–5 | 1121 | 4.56 | 12.11 | 3.56 | 0.01 | 2.40 | 12.75 | 5.80 | 0.07 |
| Jan 5–6 | 1118 | 1.96 | 6.81 | 2.31 | 0.06 | 1.22 | 10.43 | 5.88 | 0.37 |
| Jan 6–7 | 1117 | 5.00 | 10.15 | 3.10 | 0.04 | 2.13 | 12.83 | 5.87 | 0.34 |
| Total  | 6804  | 4.29 | 11.16 | 3.60 | 0.06 | 2.51 | 12.29 | 5.21 | 0.16 |
| SR    |         |    |      |      |      |     |    |      |      |      |     |     |
| Jan 1–2 | 1165 | 1.15 | 5.28 | 2.75 | 0.44 | 1.26 | 5.22 | 2.70 | 0.46 |
| Jan 2–3 | 1163 | 1.00 | 5.85 | 3.04 | 0.55 | 0.86 | 6.45 | 3.06 | 0.52 |
| Jan 3–4 | 1120 | 0.80 | 5.50 | 2.59 | 0.58 | 0.72 | 5.36 | 2.97 | 0.59 |
| Jan 4–5 | 1121 | 1.17 | 6.13 | 2.67 | 0.36 | 0.70 | 6.86 | 3.43 | 0.62 |
| Jan 5–6 | 1118 | 1.82 | 6.57 | 3.04 | 0.20 | 0.72 | 8.01 | 3.81 | 0.63 |
| Jan 6–7 | 1117 | 1.65 | 5.83 | 2.84 | 0.39 | 0.69 | 7.31 | 3.54 | 0.69 |
| Total  | 6804  | 1.19 | 5.87 | 2.82 | 0.43 | 0.76 | 6.75 | 3.24 | 0.61 |

The table above shows the quantitative evaluation of the NN and SR drift vectors compared with the vectors from the original AMSR2 images, for all image pairs individually and aggregated. "δx" and "δy" indicate the drifts on the vertical and horizontal axes, respectively. "Samples" stands for the total number of drift vectors compared, "RSE" for the relative squared error, "RMSE" the root mean squared error in km, "MAE" the mean absolute error in km, and "P" the Pearson correlation coefficient.
Figure 3.13: Error distributions from the (a) NN and (b) SR drift estimates in the two axes compared with the vectors from the original images considered as the ground-truth, for the image pairs on January 1 and 2. The histograms of the probability distribution functions appear on the bottom and right part of the plot, for the \( x \)- and \( y \)-axis drifts, respectively.

<table>
<thead>
<tr>
<th>Dates</th>
<th>( \mu_x )</th>
<th>( \mu_y )</th>
<th>( \sigma_x )</th>
<th>( \sigma_y )</th>
<th>( \mu_x )</th>
<th>( \mu_y )</th>
<th>( \sigma_x )</th>
<th>( \sigma_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 1–2</td>
<td>1.03</td>
<td>0.58</td>
<td>12.19</td>
<td>12.27</td>
<td>0.30</td>
<td>0.57</td>
<td>5.28</td>
<td>5.19</td>
</tr>
<tr>
<td>Jan 2–3</td>
<td>0.53</td>
<td>1.59</td>
<td>13.13</td>
<td>12.64</td>
<td>-0.67</td>
<td>1.02</td>
<td>5.82</td>
<td>6.37</td>
</tr>
<tr>
<td>Jan 3–4</td>
<td>1.05</td>
<td>1.36</td>
<td>11.21</td>
<td>12.48</td>
<td>0.23</td>
<td>0.60</td>
<td>5.50</td>
<td>6.33</td>
</tr>
<tr>
<td>Jan 4–5</td>
<td>1.05</td>
<td>1.31</td>
<td>12.07</td>
<td>12.69</td>
<td>0.23</td>
<td>0.11</td>
<td>6.13</td>
<td>6.86</td>
</tr>
<tr>
<td>Jan 5–6</td>
<td>0.00</td>
<td>2.09</td>
<td>6.81</td>
<td>10.22</td>
<td>-0.12</td>
<td>0.68</td>
<td>6.57</td>
<td>7.99</td>
</tr>
<tr>
<td>Jan 6–7</td>
<td>0.64</td>
<td>0.89</td>
<td>10.13</td>
<td>12.81</td>
<td>0.09</td>
<td>0.24</td>
<td>5.83</td>
<td>7.31</td>
</tr>
<tr>
<td>Total</td>
<td>0.72</td>
<td>1.30</td>
<td>11.14</td>
<td>12.23</td>
<td>0.01</td>
<td>0.54</td>
<td>5.87</td>
<td>6.73</td>
</tr>
</tbody>
</table>

Table 3.3: Quantitative evaluation of the errors of the NN and SR drift vectors compared with the vectors from the original AMSR2 images, for all image pairs individually and aggregated. “\( \mu_x \)” and “\( \mu_y \)” stand for the mean error values in the \( x \) and \( y \) directions, respectively, whereas “\( \sigma_x \)” and “\( \sigma_y \)” for the respective standard deviation.
CHAPTER 3. HYBRID EXAMPLE-BASED SUPER-RESOLUTION

Object Tracking

Similar to the previous experiments, we apply context tracker [23] to the low-resolution input frames and the super-resolved frames for comparison. By comparing the tracking results with the ground-truth results, a boost in the tracking performance is observed after super-resolution.

Fig. 3.14 illustrates the tracking comparisons over the selected frames of the whole sea ice movement period. Three different ice fragments are tracked as marked in bounding boxes with different colors. The bounding boxes are manually labeled and are the same in the starting frame. As observed, starting from frame 12 (i.e., on Feb. 28), context tracker is unable to locate one of the ice fragment instance (marked in red) within the low-resolution input. However, on the contrary, the tracking results over the corresponding super-resolved frame are almost identical compared with the ground-truth. As time goes by, context tracker failures keep occurring in the low-resolution inputs for the rest ice fragment instances. On the other hand, the tracking performance over the super-resolved imagery is stable and accurate.

The success plots for the NN and SR images are calculated to quantitatively evaluate their tracking performance compared with the tracking results from the original images, as a widely employed evaluation measure in object tracking [129]. For each image and object, the overlap score, or intersection-to-union ratio, between the region of the object bounding box \((r_t)\), and the respective ground-truth bounding box from the original image \((r_o)\), is calculated as \(d = |r_t \cap r_o|/|r_t \cup r_o|\), where the nominator and denominator represent the intersection and union between the two regions, respectively. The tracking is considered correct if \(d\) is larger than a pre-defined threshold, \(d_0\). For each of the three objects, the ratio of the number of frames (images) where the object is correctly detected to the overall number of frames is calculated. The average ratio over the three objects is the success rate, \(S\), for the specific type of image and overlap threshold. To make the evaluation more robust, we calculate \(S\)
Figure 3.14: Tracking comparisons of sea ice fragments around the Beaufort Sea spanning from Feb. 17 to Mar. 18 in year 2013. The rows represent the tracking results on the ground-truth images (Orig.), the low-resolution input images (NN), and the super-resolved images (SR), respectively, for 4 indicative dates. Three ice fragment instances are tracked in the sequential frames as marked in bounding boxes with different colors. Context tracker [23] loses track of the tracking instances in low-resolution instances but keeps a stable performance over the super-resolved frames.
Figure 3.15: Success plots of the three-object tracking with the NN and SR Suomi NPP images compared with the results from the original images, for different overlap thresholds. The Area Under Curve (AUC) values are also provided.

for different values of $d_0$ ranging from 0 to 1 (with a stride of 0.05). The Area Under Curve (AUC) is also calculated for the NN and SR plots, as a further quantitative measure of their agreement with the ground-truth results. The generated plots are provided in Fig. 3.15. As expected, $S$ generally decreases as $d$ increases, since the requirement for a correct matching gets stricter. As readily seen, the success rate for the SR images is significantly higher than the NN images, for almost the entire range of overlap thresholds. In particular, $S$ for the NN images drops by around 20% immediately after $d_0 > 0$ is applied. On the contrary, $S$ remains above 0.8 for the SR images until $d_0 = 0.4$. The AUC value for NN is 0.40, whereas for SR it is 0.61, more than 50% higher.
3.3 Discussions

In this Chapter, a hybrid example-based super-resolution framework is proposed to benefit from both external statistics (i.e., stable super-resolution performance without patch-wise similarity comparison during the online super-resolution) and internal statistics (i.e., reconstruction of structural edges and fine textures using self patch redundancy). Compared with the super-resolution framework presented in Chapter 2, it is faster and with better super-resolution performance. The use of external statistics in this framework does not conflict with the essential goal of this dissertation, i.e., to fully explore the internal statistics. External dataset is used to further boost the performance, speedup the process, and break the limitations of internal statistics.

We also successfully link the super-resolution algorithm design with a real-world application by presenting a sequential fine scale sea ice motion tracking framework. The tracking performance over super-resolved imagery significantly outperforms the one over the low-resolution data. The transferability of the proposed framework is further demonstrated in a different setting by applying a state-of-the-art object tracking algorithm in the super-resolved imagery. An obvious boost in the tracking performance is observed. In the future work, we aim to apply the designed image enhancement algorithms in assisting more real-world applications.
Chapter 4

Resolution Enhancement in Single Depth Map and Aligned Image

Depth maps are convenient in representing and storing the distance information of the objects’ surfaces given a viewpoint. They can be easily obtained through 3D imaging hardware such as TOF cameras and cost-effective consumer RGB-D cameras (e.g., Microsoft Kinect camera). As mentioned in Sec. 1.2, quality of the captured depth maps are crucial in their relevant applications, e.g., reliable 3D reconstruction, accurate human pose recognition, proper semantic scene analysis, and other geometry-related computer vision systems. However, due to the limitations of the depth sensors, depth maps suffer from low spatial resolution especially when the objects are far from the camera. Moreover, missing depth values exist due to the short distance between the object and the depth camera, disparity between the projector and the sensor, or poor reflection of the light patterns [105]. Under these circumstances, we reply on computer vision algorithms to enhance the quality of the depth maps.

Depth maps can be viewed as grayscale images where each pixel stores the depth information. Different from research in intensity image super-resolution, single depth super-resolution is not that commonly seen. From the input aspect, research in depth image res-
olution enhancement mainly falls into two classes: multi-frame super-resolution and depth image super-resolution with the assistance of an aligned high-resolution RGB image. Multi-frame super-resolution methods make use of the presence of aliasing in multiple depth inputs of the same scene to produce one fine-resolution depth map. Schuon et al. [100] verified that multi-frame super-resolution designed for intensity images also function in the 3D domain. They applied image super-resolution scheme proposed in [30] to depth images taken by a 3DV™ TOF Camera. Campbell et al. [13] adopted a discrete label MRF optimization to pose a spatial consistency constraint in extracting the fine depth map based on stored depth hypotheses. In [101], Schuon et al. incorporated a data fidelity term and a geometry prior term into an optimization framework. The former constraint ensures the fidelity between the high-resolution depth map and the low-resolution measurements and the latter term guides the energy minimization to a plausible solution. Later, Cui et al. [19] proposed a probabilistic scan alignment approach to fuse noisy scans into high quality 3D shapes. Real-time GPU-based algorithm was designed in [51] to merge low-resolution images captured by Kinect and accomplish 3D reconstruction.

With the presence of an aligned high-resolution RGB image, the second category of depth image super-resolution tends to jointly use both depth and color information of the same scene. Statistical dependency exists between the registered intensity and depth images based on the observation that depth discontinuities often co-occur with intensity changes. In [143], Yang et al. utilized the high-resolution intensity image to build the cost volume and iteratively refined the input low-resolution range image. Park et al. [92] introduced nonlocal means filtering to regularize depth maps during the reconstruction and the high-resolution intensity input provides additional features to better preserve the structure. Li et al. [72] utilized piece-wise planar assumption to regulate global geometry of the scene and proposed a Bayesian approach by taking the uncertainty of depth measurements into consideration. Kiechle et al. [56] presented a bimodal co-sparse analysis model to capture the
CHAPTER 4. ENHANCEMENT IN DEPTH MAP AND ALIGNED IMAGE

interdependency of registered intensity and depth information. In [71], a unified framework is proposed to combine multiple constrains where the aligned high-resolution intensity image could be incorporated as an additional term if available. Moreover, to obtain high-quality depth maps, missing values in the input depth map need to be filled using image completion techniques. Background information and related work can be found in Sec. 1.2.

Among numerous work in depth image enhancement, little attention has been paid to the situation where only one pair of registered low-resolution RGB image and depth map is available. Lee and Lee [68] employed a convex optimization framework for simultaneous estimation of super-resolved depth map and intensity image but required low-resolution depth map sequences as inputs. In this chapter\(^1\), we propose a novel sequential resolution enhancement framework which takes only one pair of aligned low-resolution depth and intensity images as input to obtain both high-resolution intensity image and depth map. By exploiting the statistical dependency between the input pair, a label matrix is learned through a support vector machine (SVM) classifier to differentiate the foreground objects and the background scene. Guided by the aligned low-resolution RGB image and the constructed label matrix, the missing values in the low-resolution depth map are predicted. Afterwards, the high-resolution depth map and intensity image are recovered through a set of pre-built regression models learned from external exemplars. Fig. 4.1 presents an example of the high-resolution depth map and intensity image generated by the proposed framework over “cone” in dataset [99] under the magnification factor of 4 (due to the space limit, in order to illustrate details more clearly, only part of the image is presented). As observed, missing pixels in the input depth map are correctly predicted consistent with the structure revealed by the registered intensity image. After resolution enhancement, the high-resolution depth map and intensity image have finer details.

\(^1\)This chapter was previously published in [132].
Figure 4.1: Enhancement results of “cone” (partial) (×4) in dataset [99]. Left column presents the low-resolution depth map and the intensity image and the right column shows the corresponding high-resolution results generated by the proposed enhancement framework. The low-resolution instances are upsampled by nearest-neighbor interpolation for a better illustration.
4.1 Framework Details

Provided a pair of registered low-resolution depth map and RGB image, we aim at recovering their fine-resolution correspondences with missing depth values predicted. The depth completion and super-resolution are performed in a sequential manner. A completed low-resolution depth map is first recovered in which all the missing depth pixels are predicted guided by the aligned intensity image. Traditional completion algorithms focus on reconstructing visually plausible images. However, in depth domain, an accurate prediction consistent with the intensity image is also important. Therefore, in the proposed framework, a label map which differentiates the foreground objects from the background scene is generated through a SVM classifier. The input of the SVM classifier is a pixel-level feature which encodes the local color and texture characteristics extracted from the intensity image. The training samples of the classifier are labeled using the available depth information. Afterwards, with the assistance of the label map, a guided completion is performed to predict the missing depth values. Finally, the high-resolution depth map and RGB image are recovered by feeding the completed depth map and low-resolution intensity image into a group of pre-learned regression models.

4.1.1 Depth Completion

Given an input depth map $D_l$ and its aligned low-resolution RGB image $I_l$, we first recover a completed depth map $D_{lc}$ where the missing pixel values in $D_l$ are predicted. Fig. 4.2 presents the flowchart of the depth completion process. In order to preserve the correct structure during the completion, a label map is first generated utilizing a SVM classifier.

The pixel-level features are extracted using Gabor filters and local homogeneity model to encode both color and texture information. Gabor filters have been successful in a variety of image processing related applications including image segmentation [52, 126] and texture
Figure 4.2: Schematic flowchart of the proposed depth completion process. Pixel-level features are extracted based on the intensity image. The extracted features serve as the inputs to a SVM classifier which is learned via the training data labeled by the known values in the input depth map. A guided completion is then performed over the input depth map with the assistance of the label map generated by the SVM model and the aligned RGB image.

classification \([10, 40]\). In the spatial domain, a 2D Gabor filter is a 2D Fourier basis function multiplied by an origin-centered Gaussian function, defined as:

\[
G(x, y) = \exp\left(-\frac{x^2 + y^2}{\sigma^2}\right) \exp(2\pi \theta i(x \cos \phi + y \sin \phi)) \tag{4.1}
\]

where \(\theta\) represents the spatial frequency, \(\phi\) stands for the corresponding orientation, \(\sigma\) indicates the standard deviation of the Gaussian kernel.

Gabor features are constructed from responses of Gabor filters by utilizing multiple filters on different frequencies and orientations. In our implementation, we apply the Gabor filters to the luminance channel of the color image. Filters of 5 scales at 8 orientations are adopted and the complex responses are expanded into real and imaginary parts, respectively.

The second part of the pixel-wise feature is extracted according to the local homogeneity model. The color image is first transformed from RGB to CIE Lab color space. For each component, we extract pixel-level feature which encodes the local intensity information. Based on component \(i\) \((i \in L, a, b)\), matrices \(g^i\) and \(d^i\) are computed where \(g^i\) stands for the gradient magnitude and \(d^i\) represents the standard deviation for each pixel within a
a \times a$ neighborhood centered at it. $g^i$ and $d^i$ are then normalized to range between 0 and 1. The normalized matrices are denoted as $\overline{g}^i$ and $\overline{d}^i$. For pixel $j$ under component $i$, the extracted feature $f^i_j = 1 - \overline{d}^i_j \cdot \overline{g}^i_j$ in which $\cdot$ represents the dot product of two matrices. The final feature is represented as the concatenation of the features extracted from the three components along with the Gabor responses.

The extracted features are employed as inputs to a SVM classifier. To train the SVM model, we first label a portion of the pixels utilizing the available depth information. Generally speaking, the foreground objects are closer to the camera and the background scene has larger depth values. Therefore, after eliminating the pixels whose depth values are unavailable, we annotate the $N_1$ pixels with the largest depth values with label $l_1$ and the $N_2$ pixels with the smallest depth values with label $l_2$. The annotated data serves as the training samples for the SVM model. We then apply the trained SVM classifier to predict the labels of all the remaining pixels including those whose depth values are unknown. The labelling results including the previous training labels are saved to a label map $L_d$ which later is utilized to assist the depth completion process.

We denote the unknown region(s) in $D_l$ as mask $M$. The completion order of the masked depth pixels is calculated using Eq. 2.4. After computing the priority for every pixel along the boundary of the masked region, the depth value of pixel $p$ with the highest priority is predicted first. We calculate the similarity between pixel $p$ with the qualified pixel(s) within a neighborhood of $b \times b$ centered at $p$. Only pixels with known depth values and share the same label as $p$ measured in the label map are considered qualified. The similarity between two pixels is measured in the mean square error of the corresponding pixel-level features introduced above. Then the depth value for pixel $p$ is filled with the corresponding depth value of the most similar one. After updating the confidence term and the data term, the above process is repeated until all the depth values within mask $M$ are predicted.
CHAPTER 4. ENHANCEMENT IN DEPTH MAP AND ALIGNED IMAGE

4.1.2 Super Resolution

As aforementioned, single image super-resolution is a numerically ill-posed problem and therefore relies on additional assumptions or priors to finalize the output. In this chapter, depth map super-resolution is investigated, which is different from the traditional super-resolution task for natural RGB images. One major difference lies in the fact that depth maps in general do not have complicated edges or texture patterns.

In the proposed super-resolution framework, we adopt a modified version of the hybrid example-based super-resolution framework introduced in Sec. 3.1. Specifically, only the ’proxy image recovery’ step is employed. The completed depth map and the input intensity image are fed into a group of externally-trained regression models to recover the final high-resolution outputs. The regression models are trained separately for depth maps and intensity images utilizing the same learning pipeline over different training datasets.

Fig. 4.3 illustrates the schematic pipeline of the super-resolution process. A large set of high-resolution/low-resolution exemplar patch pairs with magnification factor $s$ are collected from a dataset consisted of training images. Original images in the dataset are considered high-resolution images and the corresponding low-resolution images are generated through
a blur and downsampling process. For an instance patch pair, both patches are normalized by extracting the mean value of the low-resolution patch.

Same as the hybrid example-based super-resolution, we first model the input feature space with GMM of $K$ components to ensure a more targeted learning (refer to Eqs. 3.1 and 3.2 for more details). We then assign each low-resolution feature to the corresponding Gaussian component according to the posterior. Afterwards, a linear regression model is trained within each Gaussian component using Eq. 3.3.

After performing the training process, $K$ regression models are learned for upscaling the intensity images (‘gradient-level self-awareness’ could be performed afterwards if necessary. Details can be found in Sec. 3.1.1). Another set of $K$ regression models is obtained through learning over a depth map dataset in a similar manner. Given a testing low-resolution instance, features are extracted by performing normalization and vectorization for each patch. After that, according to the posterior, each feature is assigned to a Gaussian component where the corresponding regression model is applied to obtain the high-resolution instance. Weighted average is adopted to blend overlapping pixels. Back-projection is utilized as post processing.

4.2 Experimental Results

In this section, the proposed resolution enhancement system is evaluated on the Middlebury Stereo Datasets [20, 46, 98, 99] and RGBD Scenes dataset v2 [64]. We present multiple results and compare them with the recent state-of-the-art approaches both quantitatively and qualitatively.
Table 4.1: Comparison of the proposed approach with other depth super-resolution schemes over 14 depth maps in dataset [99] measured in terms of FSIM [146] under the magnification factor of 4. The best performance in FSIM for each image is marked in bold.

4.2.1 Implementation Details

During the depth completion, the pixel-level feature vector has a dimension of 83 in which 80 of them come from the Gabor responses of multiple filters at 5 scales and 8 orientations. The rest are extracted according to the local homogeneity model. The neighborhood size \( a \) defined to calculate the standard deviation is set to 5. The numbers of training samples \( N_1 \) and \( N_2 \) are the same and equal to 10% of the total number of pixels in the low-resolution RGB image. We employ SVM with RBF kernel utilizing LIBSVM [15].

In the super-resolution phase, the training dataset used for intensity regression model learning is the same as in Sec. 3.1 with 6,152 natural images. 1,449 completed depth maps in NYU Depth Dataset V2 [106] are employed to train the depth regression models. Parameter settings are the same as shown in Sec. 3.1.2. The number of iterations for back-projection is set to 10. The time cost of the proposed system varies depending on the size of the input image, the scaling factor, and the number of missing pixels in the depth map. Generally speaking, under a scaling factor of 4, for an input depth map \((120 \times 100)\) with missing pixels less than 10% of the overall number of pixels, it takes several seconds to finish the depth completion and the super-resolution.
### Table 4.2: Comparison of the proposed approach with peer methods for intensity images in Middlebury Stereo datasets [20, 46, 98, 99] measured in terms of average FSIM [146] under the scaling factor of 4. The best performance in FSIM for each dataset is marked in bold.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M2001 [99]</td>
<td>0.7034</td>
<td>0.7705</td>
<td>0.7716</td>
<td>0.8052</td>
<td><strong>0.8169</strong></td>
<td>0.8129</td>
</tr>
<tr>
<td>M2003 [20]</td>
<td>0.7128</td>
<td>0.7894</td>
<td>0.7932</td>
<td>0.8200</td>
<td>0.8213</td>
<td><strong>0.8215</strong></td>
</tr>
<tr>
<td>M2005 [98]</td>
<td>0.9826</td>
<td>0.9863</td>
<td>0.9702</td>
<td>0.9930</td>
<td>0.9902</td>
<td><strong>0.9940</strong></td>
</tr>
<tr>
<td>M2006 [46]</td>
<td>0.9855</td>
<td>0.9892</td>
<td>0.9728</td>
<td>0.9944</td>
<td>0.9918</td>
<td><strong>0.9949</strong></td>
</tr>
</tbody>
</table>

### 4.2.2 Quantitative Comparison

We evaluate the proposed enhancement framework on a variety of depth maps and RGB images in Middlebury Stereo Datasets [20, 46, 98, 99] under the scaling factor 4. Middlebury dataset 2001 [99] consists of depth maps without missing values and depth maps provided in the rest datasets, i.e., Middlebury datasets 2003, 2005, 2006, all contain unknown regions. In our experiments, the low-resolution inputs are generated from the depth maps and intensity images in the datasets by performing the nearest-neighbor downsampling. Since the completed ground-truth depth maps for Middlebury datasets 2003, 2005, 2006 are unavailable, it is infeasible to provide numerical statistics to measure the depth completion performance.

Quantitative evaluations for super-resolution performance are provided over depth maps in Middlebury dataset 2001 and intensity images for the all listed datasets. The recent proposed image quality assessment criterion Feature Similarity (FSIM) [146] is adopted for measurement since the human visual system understands an image mainly according to its low-level features and therefore FSIM achieves higher consistency with the subjective evaluation compared with other metrics. In the proposed depth super-resolution, statistics from the aligned RGB image is not utilized. Therefore, we compare our results with the state-of-the-art single depth super-resolution method [4]. As illustrated in Table 4.1, our approach outperforms [4] in 13 out of 14 depth maps measured in terms of FSIM. [120, 139, 141] are the state-of-the-art external example-based single image super-resolution approaches.
5.2 presents the comparison of the proposed approach with these methods listed for the intensity images in 4 different datasets. Our framework adopts GMM to model the input feature space and therefore a more targeted learning is ensured for the regression model training. Statistics from Table 4.2 reveals that the proposed super-resolution framework is the most effective one in 3 out of 4 evaluated datasets for intensity outputs measured in average FSIM.

### 4.2.3 Qualitative Comparison

The enhancement performance is further evaluated qualitatively. Fig. 4.4 presents our enhancement results under the scaling factor 4 of “Baby3”, “Midd1”, “Moebius”, and “Cone” from the Middlebury datasets. In “Baby3”, it is challenging to correctly predict the missing depth values due to the cluttered background (i.e., map with irregular contours and patterns) and the color similarity between certain foreground objects and the background scene. As observed, after the proposed enhancement framework, missing depth values are well filled in a manner consistent with the embedded structure. Effectiveness of the proposed depth completion algorithm can be further demonstrated in “Cone” where complex texture patterns exist along with large missing depth areas. Fine details and clear edges are preserved as illustrated from both the high-resolution depth maps and the high-resolution intensity outputs.

Fig. 4.5 presents more enhancement results in RGBD Scenes dataset v2 [64] under the magnification factor of 4. Different from the Middlebury datasets, depth maps in [64] suffer from large missing areas and the registered intensity images are of low quality with blurry visual artifacts. The proposed enhancement framework manages to reconstruct structurally correct completed depth maps. Clear contours are recovered in the high-resolution results.

We further compare our results with representative peer algorithms [4, 120, 141] in depth and intensity super-resolution. In Fig. 4.6, a set of super-resolution results is provided on
Figure 4.4: Resolution enhancement results of “Baby3”, “Midd1”, “Moebius”, and “Cone” (×4). From left to right, the columns represent low-resolution input depth maps, generated high-resolution depth maps, low-resolution intensity images, and generated high-resolution intensity outputs. For a better presentation, the input instances are upsampled through nearest-neighbor interpolation.
Figure 4.5: Resolution enhancement results in RGBD Scenes dataset v2 [64] (×4). From left to right, the columns represent low-resolution input depth maps, generated high-resolution depth maps, low-resolution intensity images, and generated high-resolution intensity outputs. For a better presentation, the input instances are upsampled through nearest-neighbor interpolation.

Figure 4.6: Super-resolution results of depth maps “bull” and “sawtooth” (×4). From left to right, the columns represent the ground-truth, results generated utilizing nearest-neighbor interpolation, patch-based [4], and the proposed framework. Only partial of the original depth maps are shown for a clearer presentation.
Figure 4.7: Super-resolution results of RGB images “bull” and “sawtooth” (×4). From left to right, the columns represent the results generated through nearest-neighbor interpolation, bicubic interpolation, ScSR [141], GR [120], and the proposed system. Only partial of the images are shown for a clearer presentation.

depth maps “bull” and “sawtooth” compared with nearest-neighbor interpolation and the outputs generated by [4]. For a better illustration, only part of each depth map is presented. The nearest-neighbor interpolated results suffer from blurry visual artifacts. Irregular zigzag patterns occur in results constructed by [4]. Compared with the ground-truths, our results best recover the contours with minimal visual artifacts. Corresponding high-resolution intensity outputs are presented in Fig. 4.7 compared with the nearest-neighbor interpolation, bicubic interpolation, ScSR [141], and GR [120]. While interpolation-based methods produce over-smoothed results, ringing artifact exists in results generated by [120]. [141] reconstructs results with gridded patterns that do not exist in the original image. As observed from the edges and textures, our recovered high-resolution images reveal more natural patterns and finer details.

4.3 Discussions

In this Chapter, we extend our research to the depth domain and accomplish the depth map completion and super-resolution. A novel framework is proposed to handle the situation
where only one pair of registered low-resolution RGB image and depth map is available. The internal statistical dependency between the coupled pair is fully explored and utilized to extract the features, train the SVM classifier, and generate the label map to assist the depth completion process. For the depth map super-resolution, since in general depth maps do not contain complicated textures, external example-based super-resolution is applied for the maximum efficiency.
Chapter 5

Self-Guiding Multimodal LSTM -
when we do not have a perfect
training dataset for image captioning

An image captioning framework is presented in this Chapter. In this study, we work with image-sentence data collected directly from Flickr. The descriptions provided by the original users are utilized as the training data. Images from Flickr have been widely used in the dataset collection [48, 88, 144]. However, the descriptions provided by the users are rarely directly used for captioning purpose due to several characteristics of the Flickr text data:
1) Lengths of the descriptions vary dramatically for each image. While some users talk in paragraphs about the details including the possible background that is not directly related to the image, others may just describe in a few words indicating the location or the date information. 2) Users may input descriptions for an album instead of a photo. Therefore, we may have multiple images that are visually different but with the same description. 3) Different from the labelling process performed by AMT workers, the content of the descriptions is not strictly controlled semantically or syntactically. Foreign languages exist along
with personal information including copyright statement, camera information, and links to personal social media accounts. Existing natural language processing tools provide a limited solution in preparing the training data. In the end, it becomes tedious to set filtering criteria or use regular expressions to generate the ‘perfect’ training dataset.

However, despite all the characteristics listed, Flickr data meets the criterion for image captioning task – it comes from millions of users who can describe anything related to the images they upload. And more importantly, it is a real-world valuable resource. In this section, we use ‘new york city’ as our test case, i.e., ‘new york city’ is employed as the keyword for the query process, to build FlickrNYC dataset. We observe in FlickrNYC that, descriptions in shorter lengths are more strongly correlated to the image content and are mainly related to the locations, events, or activities. They occur more repetitively compared with longer descriptions, e.g., a user may uploaded several images related to a walk in central park and they all have the description as ‘central park’. On the other hand, descriptions in longer sentences or paragraphs reveal more syntactical details, but may provide concepts that are more implicitly related to the images and have a weaker or no correlation to the image content. Examples images of FlickrNYC can be found in Fig. 5.1.

In the proposed framework, a self-guiding multimodal long short-term memory (sg-LSTM) framework is presented to leverage between two portions of the data: $data_s$ (images with shorter length of descriptions) and $data_l$ (images with longer length of descriptions). We aim to make use of the part of the dataset with more reliable information to guide the training process of the caption generation. As demonstrated in Fig. 5.2, a direct training utilizing the state-of-the-art multimodal RNN captioning method [77] fails to capture the core event revealed in the image due to the fact that, FlickrNYC is a noisy real-world dataset in which we have multiple images labelled as ‘thanksgiving’ but are visually different. Moreover, thanksgiving celebration is more frequently seen than Chinese New Year celebration in the training dataset. However, the proposed framework manages to generate accurate
Figure 5.1: Sample images and descriptions in FlickrNYC dataset. (a) Examples from $data_s$ in which images are with short descriptions. (b) Examples from $data_l$ in which images are with long descriptions.
Figure 5.2: Example of the description generated by the proposed sg-LSTM image captioning framework compared with the result output by the traditional multimodal RNN. Both frameworks are trained with FlickrNYC – our newly proposed dataset.

description that is both semantically and syntactically correct.

5.1 Framework Details

The successful combination of CNN and RNN, especially LSTM, has been widely experimented in image captioning and related tasks. However, as observed in [53], the generated sentence is sometimes weakly coupled to the provided image but is strongly correlated to the high frequency sentences in the training dataset. This is due to the fact that the generated sentence is “drifted away” during the sequence prediction process. This problem exists especially for long sentences where the generation is carried out “almost blindly towards the end of the sentence”. To address this issue, alternative extensions have been proposed by adding attention mechanism [135] and modifying the LSTM cell [53]. However, it is still
CHAPTER 5. SELF-GUIDING MULTIMODAL LSTM

Figure 5.3: Systematic flowchart of the proposed sg-LSTM captioning framework. (a) Basic multimodal LSTM (m-LSTM) captioning framework trained on a subset of the FlickrNYC dataset with short descriptions (i.e., \( data_s \)). \( w_t \) denotes the \( t \)-th word in a sentence with words ranging from \( w_1 \) to \( w_T \). A start sign \( w_{\text{start}} \) and an end sign \( w_{\text{end}} \) are added to all training instances in both \( data_s \) and \( data_l \). (b) Guiding text feature (GTF) extraction: to extract the text feature for self-guiding, we generate the descriptions utilizing m-LSTM followed by a sentence vectorizer. (c) Illustration of the sg-LSTM architecture. Compared with m-LSTM, an additional textual feature is fed to the multimodal block which encodes the language information connected to the image content.
challenging to work on an uncontrolled dataset with descriptions in arbitrary lengths and different abstraction levels.

In this section, we first introduce the basic multimodal LSTM (m-LSTM) image captioning framework which fuses the information of the input sentences and the corresponding image features in the multimodal component. It works effectively when the two input sources are strongly bonded. However, when this is not the case, it is difficult to maintain the correlation as the sentence generation goes on especially when the training dataset is not ideal for image captioning task.

As mentioned, in FlickrNYC dataset, descriptions in shorter sentences tend to have a stronger bond with the image content compared with longer descriptions. Although they may not be syntactically sound to form a sentence, these short descriptions tend to accurately describe the locations, activities, objects, or events, as the images were taken. Some examples can be found in Fig. 5.1(a) where core information in these images are conveyed in the corresponding descriptions. On the other hand, long descriptions are valuable as the users may state their feelings, reasoning, personal experiences, or objects that are not depicted in the images. As shown in Fig. 5.1(b), these sentences are difficult to reproduce by the AMT workers even with specific instructions. However, some descriptions may not be strongly bonded with the visual content.

In order to generate image captions with adequate details related to the image content, we separate the data based on the different characteristics revealed. FlickNYC is divided into two subsets, \( data_s \) with descriptions in short sentences or terms, and \( data_l \) with descriptions in long sentences or paragraphs (the length is measured in the number of words). We start by training a m-LSTM captioning model based on \( data_s \). This captioning model aims to extract the key textual information provided an input image. This key information is later utilized to guide the training of sg-LSTM based on \( data_l \) to better link the description with the image content. The guiding information is represented using a sentence vectorizer and
fed as another input to the multimodal component in sg-LSTM.

## 5.1.1 Captioning with m-LSTM

To train a caption model with data, we employ a variation of m-RNN due to its elegance and simplicity. The gated recurrent unit is replaced with LSTM in the proposed m-LSTM model. The LSTM network has been widely used to model temporal dynamics in sequences. Compared with the traditional RNN, it better addresses the issue of exploding and vanishing gradients. The basic LSTM block consists of a memory cell which stores the state over time and the gates which control how to update the state of the cell.

As illustrated in Fig. 5.3(a), m-LSTM is composed of a word embedding layer, an LSTM layer, a multimodal layer, and a softmax layer. It takes the training images and the corresponding descriptions as inputs. Each word in the sentences is encoded with one-hot representation before being fed to m-LSTM training. The word embedding layer aims to map the one-hot vector to a more compact representation as shown in Eq. 5.1. Same as [77], we randomly initialize the embedding layer and learn $W_e$ during training.

$$e_t = W_e \cdot w_t$$  \hspace{1cm} (5.1)

where $w_t$ stands for the one-hot representation of word at step $t$. $W_e$ is the mapping weight between the one-hot representation and the word embedding representation $e_t$.

There are many LSTM variants. In the proposed m-LSTM model and later in sg-LSTM, we adopt LSTM with peepholes where the memory cell and gates within an LSTM block are defined as:

$$i_t = \sigma(W_{ic}C_{t-1} + W_{ih}h_{t-1} + W_{ie}e_t + b_i)$$  \hspace{1cm} (5.2)

$$f_t = \sigma(W_{fc}C_{t-1} + W_{fh}h_{t-1} + W_{fe}e_t + b_f)$$  \hspace{1cm} (5.3)
in which \( \odot \) denotes the element-wise product. \( \sigma(\cdot) \) is the sigmoid nonlinearity-introduce function. \( g_1(\cdot) \) is the basic hyperbolic tangent function. \( i_t, o_t, f_t, C_t, \) and \( h_t \) represent the state values of the input gate, output gate, forget gate, cell state, and hidden state, respectively. \( W_{ci} \) and \( b_i \) denote the weight matrices and bias vectors for corresponding gates and states.

The word sequence is fed to the LSTM network by iterating the recurrence connection as shown in Fig. 5.3(a). Inception v3 [114] is used to extract the image features. They are connected with the language inputs through a multimodal component. The multimodal block fuses the language information represented as the dense word embedding and the LSTM activation with the image information represented using CNN as shown below:

\[
mm_t = g_2(W_i \cdot I + W_d \cdot e_t + W_l \cdot \text{LSTM}_t) \quad (5.7)
\]

where \( g_2(\cdot) \) is the element-wise scaled hyperbolic tangent function [67] which leads to a faster training process than the basic hyperbolic tangent function. \( W_i, W_d \) and \( W_l \) indicate the mapping weights to learn during training.

The m-LSTM model is learnt utilizing a log-likelihood cost function based on perplexity introduced in [77]:

\[
C_{m-LSTM} = \frac{1}{N_w} \sum_{i=1}^{N_s} L_i \cdot \log_2 \mathcal{P}(w^{(i)}_{1:L_i} | I^{(i)}) + \lambda_\theta \cdot \| \theta \|_2^2 \quad (5.8)
\]

where \( \mathcal{P}(\cdot) \) stands for the perplexity of a sentence given an image. \( N_w \) and \( N_s \) represent the
number of words and the number of sentences in the training set. \( L_i \) is the length of the \( i \)-th sentence, and \( \theta \) denotes the model parameters.

### 5.1.2 Captioning with sg-LSTM

In this subsection, we describe in detail the training of sg-LSTM with \( data_t \). As mentioned, for some training instances in \( data_t \), there is not a strong connection between the textual description and the image content. In other words, additional textual features are needed during training. Therefore, as presented in Fig. 5.3(b), a guiding textual feature (GTF) extractor is proposed which connects a m-LSTM captioning model trained on \( data_s \) to a sentence vectorizer. This guidance feature aims to provide additional textual information for each training instance in \( data_t \), which emphasizes the correlation between the textual and the visual domains. Compared with the basic m-LSTM architecture, sg-LSTM carries additional information in the multimodal component. Same as the image feature, the guiding textual features are fed into the multimodal component on each timestep as auxiliary information. This additional textual feature implicitly encodes the semantic information related to the image, such as location, activity, etc.

The sg-LSTM architecture is composed of four layers in each timestep similar to m-LSTM. The embedding layer encodes the one-hot word representation into a dense word representation. The weights in the embedding layer are learned from the training data aiming at encoding the syntactic and semantic meaning of the words. The word representation after the embedding layer serves as the input to the LSTM layer. Same as m-LSTM, we adopt a basic LSTM block with peepholes. After this layer, a multimodal layer is set to connect the CNN-based image feature, the dense word representation, the recurrent layer output, and the proposed guiding texture feature. The activations of these four inputs are mapped to the same multimodal feature space as the activation of the multimodal layer:
\[
m m_t^i = g_2(W_i^i \cdot I + W_d^i \cdot e_t + W_l^i \cdot \text{LSTM}_t + W_t^i \cdot T)
\]  
(5.9)

where \( W_i^i, W_d^i, W_l^i, \) and \( W_t^i \) represent the corresponding weighting matrices.

**Extraction of Guiding Textual Features:**

To generate the guiding textual feature for a certain image, we first utilize m-LSTM trained on data\( _s \) to output a short description (i.e., the raw sentence for the guiding textual feature). Beam search is adopted in the process to avoid the exhaustive search in the exponential search space. It is widely used in RNN-based captioning models [53, 77, 91] due to its efficiency and effectiveness. The top 1 ranked sentence is selected for further vectorization. Fig. 5.4 presents examples of the guiding texts generated by m-LSTM trained on data\( _s \). The images are from data\( _l \) and therefore, the original descriptions are relatively long. As demonstrated, the guiding texts either provide core information that is not conveyed in the original descriptions, e.g., authorship info (jackson pollock), landmark name (radio city music hall), and season info (snowy day), or emphasize the key image content buried in long sentences, e.g., event (macy’s thanksgiving day parade), and location (grand central terminal). We also observe some interesting results that reveal some underneath feelings of the images themselves, e.g., ‘snow, dirt, love, and loneliness’.

A group of sentence vectorizers are investigated to vectorize the sentence or term generated by m-LSTM. In general, we adopt the word2vec with fusion scheme, i.e., each word in the sentence is vectorized and then these word vectors are combined to produce the final output. Three word2vec schemes are experimented:

- **word2vec-GloVe**: we adopt pre-trained GloVe [94], i.e., Global Vectors for word representation, as the word vectorizer. The word vectors are trained through aggregated global word-word co-occurrence statistics from a corpus combining Wikipedia 2014 and
Figure 5.4: Examples of the guiding text (marked in red) generated by m-LSTM compared with the original descriptions (marked in blue) provided by the Flickr users. The guiding text provides supplementary information that is strongly related to the image content.
Gigaword 5. We test two different feature dimensions, 50 and 300.

- word2vec-NYC: compared with word2vec-GloVe, word2vec-NYC is a local word vectorizer trained on the textual data in FlickrNYC. This model is trained utilizing gensim [97] and a 128-dimensional vector is generated per word.

- word2vec-short: a word embedding mapping is learnt when training m-LSTM on data. In this word vectorizer, the representation after the word embedding layer is employed directly as to map a word to a 1,024-dimensional vector.

After representing each word in vector, two different fusing methods are investigated:

- Average: an average of all the word vectors in a sentence is calculated to obtain the final sentence vector.

- TF-IDF: the word vectors are combined using term frequency-inverse document frequency (TF-IDF) weighting scheme to generate the final representation.

The various vectorization methods look into the mapping problem from different angles, utilizing a global corpus or a local dataset, and in different dimensionality. As later shown in Table 5.2, sg-LSTM based on word2vec-GloVe with TF-IDF weighting under feature dimension 50 (denoted as sgLSTM-GloVe-tfidf-50) works the best among all the 8 vectorization schemes.

### Training sg-LSTM

Same as m-LSTM, a log-likelihood cost function related to the perplexity is utilized for training sg-LSTM as shown in Eq. 5.8. Normalization regarding the number of words corrects the bias over shorter sentences during the caption generation process, and therefore, is suitable for FlickrNYC with images in various lengths.
5.2 Experimental Results

In this section, the effectiveness of the proposed self-guiding strategy is verified experimentally on FlickrNYC. We start by a deeper introduction of FlickrNYC dataset followed by the implementation details of the proposed system. Afterwards, experimental evaluation results are presented and analyzed.

5.2.1 FlickrNYC Dataset

The FlickrNYC dataset is composed of 306,165 images in total collected from Flickr with keyword ‘new york city’. More specifically, Flickr search API is employed to crawl image-description data based on the keyword, i.e., photos whose title, description, or tags contain ‘new york city’ will be fetched. After capturing the images and their corresponding metadata, each image is accompanied with 1 reference description provided by the original user. Images without valid descriptions are discarded. We perform a light pre-processing utilizing NLTK Toolbox [11], textacy¹, and self-defined regular expressions, to remove unnecessary personal information (e.g., URLs, copyright declaration, camera information, personal social media accounts, advertisements, etc).

After the textual pre-processing, the dataset is divided based on the number of words in the descriptions. Images with descriptions shorter than 10 form dataset $data_s$ with 165,374 images for training and 1,000 for testing. The rest 139,791 images form $data_l$ in which 137,791 are used for training, 1,000 for validation and 1,000 for testing. Table 5.1 provides the statistics of distributions based on the description lengths in FlickrNYC. Sample images and the corresponding descriptions can be found in Fig. 5.1.

Different from the traditional way to create the vocabulary which removes all words that contain non-alphanumeric characters or even non-alphabetic characters, the vocabulary

build-up process for FlickrNYC is tricky: 1) Since the dataset is based upon New York city in which multiple landmark names contain combinations of alphanumeric characters (e.g. ‘5th avenue’ or informally ‘5 ave’ in some descriptions), therefore, numeric and alphanumeric words should not be eliminated in the vocabulary. Moreover, words that contain or are connected by punctuations should also be considered, e.g., ‘Macy’s’, ‘it’s’, ‘let’s’, ‘sightseeing’, ‘African-Americans’, etc. 2) Although one image is accompanied by one description, the description is not restricted to one sentence. As observed, some descriptions can be long containing multiple sentences. To better model the continuity of a paragraph of sentences, punctuations such as ‘’, ‘,’ ‘.’ ‘!’ and ‘?’ should be considered as part of the vocabulary list. 3) FlickrNYC utilizes uncontrolled real-world text data, which indicates that the usage of words can be informal. However, we find sometimes this informality is valuable since it reveals the emotions of the users, such as Emoticons (‘:-)’, ‘:-P’, etc.) and exaggerated expressions (‘soooo’, ‘superrrr’, etc.). Therefore, in order to keep all the information mentioned above, after tokenization and converted to lowercase, words that appear at least 3 times in the training set are kept to create the vocabulary. The final vocabulary size is 22,230.

<table>
<thead>
<tr>
<th>sentence length</th>
<th>1−5</th>
<th>6−10</th>
<th>11−15</th>
<th>16−25</th>
<th>≥26</th>
</tr>
</thead>
<tbody>
<tr>
<td>num of instances</td>
<td>94,180</td>
<td>85,691</td>
<td>45,291</td>
<td>37,108</td>
<td>43,895</td>
</tr>
</tbody>
</table>

Table 5.1: Statistics of image distribution based on the description lengths in FlickrNYC dataset.

5.2.2 Implementation Details

The proposed framework is built upon m-RNN with TensorFlow. The inception v3 pretrained on ImageNet is used to extract CNN features as image representation.

\(^2\)If a word only contains alphabetic characters, we employ WordNet to rule out typos and non-English words.

\(^3\)https://github.com/mjhucla/TF-mRNN
Table 5.2: Numerical results of the proposed framework compared with other methods based on the testing images in data_t.

<table>
<thead>
<tr>
<th>m-RNN [77]</th>
<th>m-LSTM-long</th>
<th>sgLSTM-NYC-ave</th>
<th>sgLSTM-GloVe-tfidf-50</th>
<th>sgLSTM-GloVe-tfidf-300</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>BLEU-2</td>
<td>BLEU-3</td>
<td>BLEU-4</td>
<td>METEOR</td>
</tr>
<tr>
<td>0.036</td>
<td>0.019</td>
<td>0.000</td>
<td>0.000</td>
<td>0.021</td>
</tr>
<tr>
<td>0.310</td>
<td>0.257</td>
<td>0.216</td>
<td>0.169</td>
<td>0.145</td>
</tr>
<tr>
<td>0.237</td>
<td>0.194</td>
<td>0.160</td>
<td>0.133</td>
<td>0.122</td>
</tr>
<tr>
<td><strong>0.417</strong></td>
<td><strong>0.381</strong></td>
<td><strong>0.359</strong></td>
<td><strong>0.339</strong></td>
<td><strong>0.211</strong></td>
</tr>
<tr>
<td>0.281</td>
<td>0.279</td>
<td>0.278</td>
<td>0.276</td>
<td>0.154</td>
</tr>
</tbody>
</table>

Feature dimension for this image representation is 2,048. In both m-LSTM and sg-LSTM, the word embedding layer is with 1,024 dimension. The LSTM layer and the multimodal layer are with 2,048 dimensions. We assign 0.5 dropout rate to all three layers.

Both m-LSTM and sg-LSTM models are trained with RMSProp optimizer [45]. We apply the stochastic gradient descent (SGD) with mini-batches of 64. The beam search size is set to be 3. The top ranked sentence generated by m-LSTM based on training data in data_s is utilized for guiding textual feature extraction. As mentioned, three word2vec schemes are tested, i.e., word2vec-GloVe, word2vec-NYC, and word2vec-short. Two different sets of pre-trained word vectors are tested for word2vec-GloVe with dimensions 50 and 300. Dimensions for word2vec-NYC and word2vec-short based representations are 128 and 1,024, respectively.

5.2.3 Experimental Evaluations

In order to select the best vectorization scheme for the guiding textual feature, certain objective criterion is needed to evaluate each scheme. Popular evaluation metrics for image captioning tasks include BLEU [90] (BLEU@1, 2, 3, 4), METEOR [5], ROUGE-L [73], and CIDEr [124]. However, none of the criteria listed is a perfect metric for the evaluation task in our case since the ground-truth descriptions in FlickrNYC dataset are noisy. An example can be found in the bottom left image in Fig. 5.5 in which the original description is ‘december 6th’. On the other hand, the proposed sg-LSTM framework outputs description ‘boaters on
CHAPTER 5. SELF-GUIDING MULTIMODAL LSTM

the lake in central park near the bow bridge’ which is a much better description compared with the original one given the image content. However, this superiority will not be reflected in the numerical metrics listed above.

Despite the challenges in evaluating the vectorization schemes in the proposed framework, there still exists a large portion of data in FlickrNYC which suits ‘perfectly’ for captioning task. Therefore, a small validation dataset is separated from data\_l and utilized to evaluate the 8 different vectorization schemes. Based on the experimental results, sgLSTM-GloVe-tfidf-50 achieves the best performance quantitatively. Thus we adopt sg-LSTM based on word2GloVe in TF-IDF weighting with dimension 50 as the final setting and all the results reported in this section are based on this setting unless stated otherwise.

Table 5.2 presents the numerical results based on 1,000 testing images in data\_l. The proposed sg-LSTM framework is compared with m-RNN [77], m-LSTM, and among different vectorization settings. The results of the top 3 performers in the previous verification step are included in this table. m-LSTM-long represents the m-LSTM captioning model trained on data\_l. As shown in Table 5.2, sgLSTM-GloVe-tfidf-50 gives the best performance numerically almost among all the evaluated methods, which is consistent with our observation in the verification step.

The zero numbers shown in Table 5.2 for m-RNN might be better explained if combined with the results shown in Fig. 5.5. A direct training over the whole dataset tends to put a preference into high frequency sentences in the training dataset, which may be unrelated to the test image itself. Therefore, when it comes to numerical evaluations, a total miss of the core concept in the image content leads to a low score. On the other hand, by integrating the guiding textual features into the training process, the proposed sg-LSTM model manages to generate accurate descriptions related to the image content, and sometimes, the generated descriptions are more meaningful than the original ones provided by the Flickr users as demonstrated in Fig. 5.5.
Figure 5.5: Descriptions generated by the proposed framework (marked in red) compared with m-RNN [77], m-LSTM-long (m-LSTM trained on \( data_i \)), m-LSTM-full (m-LSTM trained on all training data) and the original descriptions (marked in blue) provided by the Flickr users. The guiding texts are also provided. To help with the evaluation, the ground-truth locations are marked in each image (usage of different colors is for the best contrast).
5.3 Discussions

In this Chapter, a novel self-guiding multimodal LSTM captioning framework which targets at a more effective training over uncontrolled real-world dataset is proposed. FlickNYC dataset is introduced as the testbed to verify the proposed self-guiding scheme. The self-guiding process looks into the learning process in a global way to balance the syntactic correctness and the semantic details revealed in the images. It is worth mentioning that this scheme can be extended to handle other vision-related tasks when dealing with imperfect training data. We aim to mimic a real-world situation when we do not have the ideal environment for training or it would be too expensive to generate one. Under these circumstances, the internal statistics hidden in the training dataset should be explored and utilized to best accomplish the task.
Chapter 6

Conclusions

In this dissertation, three different tasks in vision research are investigated spanning from low-level vision to high-level vision. We target at situations where limited resources are provided, i.e., only one single image is input for image super-resolution/completion and the training dataset for image captioning task is uncontrolled real-world web data. Under these circumstances, internal statistics are fully utilized to best accomplish these tasks.

We have presented a single image super-resolution model based on internal gradient similarity. This framework can be naturally extended to handle image super-resolution and completion simultaneously. Afterwards, to speedup the super-resolution process and accomplish robust performance for large scaling factors, a hybrid example-based super-resolution algorithm is proposed which benefits from both internal and external statistics. The algorithm is applied to assist sea ice motion tracking. Later, our research extends to handle image quality enhancement for registered low-resolution RGB image and depth map where there are missing regions in the depth input. Finally, a self-guiding multimodal LSTM image captioning framework is presented to handle the uncontrolled image-sentence dataset where descriptions could be strongly or weakly correlated to the image content and in arbitrary lengths.
Appendices
Appendix A

Publications

• **Yang Xian** and YingLi Tian. Self-Guiding Multimodal LSTM - when we do not have a perfect training dataset for image captioning. Ready for submission.


• Wendy Fernandez*, **Yang Xian***, and YingLi Tian. Image-Based Barcode Detection and Recognition to Assist Visually Impaired Persons. IEEE International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (IEEE-CYBER), 2017. (*: equal contribution)


• **Yang Xian** and YingLi Tian. Resolution Enhancement in Single Depth Map and Aligned Image. IEEE Winter Conference on Applications of Computer Vision (WACV), 1-9, 2016. doi: 10.1109/WACV.2016.7477645


Bibliography


