Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computer Vision and Geometry Lab

Self-Supervised Single Image Super-Resolution for Satellite Image

Master Thesis

HUANG Yuanhao
Department of Computer Science

Advisor: Peidong Liu, Dr. Martin Ralf Oswald
Supervisor: Prof. Dr. Marc Pollefeys

May 15, 2020

Abstract

This thesis work focuses on reference-based self-supervised image super resolution problem. Recently, deep convolutional neural network has been widely adopted for image super resolution tasks due to its superior performance against traditional non-deep-learning based methods. However, majority of these proposed methods assume that ground truth is available or down-scaling is allowed to get corresponding low-resolution images where training is conducted. Few research works have been done for methods that are trained in an unsupervised way.

We propose an optical-flow-based cycle-consistent image super resolution pipeline to tackle the problem when ground truth is not available and training on the degraded scale is not allowed. This structure includes a super resolution module, an optical flow estimation module for multi-view information analysis, a warping module for image reconstruction from optical flow and references and a cycle-consistent generative adversarial network (CycleGAN)[3] module for style transformation between real high-resolution images and generated super-resolved images.

We show that generating high-quality texture features in high-resolution space through pure unsupervised learning is difficult because the network fails to learn cross-scale features efficiently. Two methods are introduced to tackle this problem: a) add external image pairs during the training of super resolution module and b) add feature extraction sharing mechanism to the super resolution module.

The dataset used for training are 284 satellite images categorized into 5 groups based on different locations. Images in the same group are taken in the same spot in different time of the year.

We present outstanding performance of our pipeline, demonstrating the improvement by adopting CycleGAN structure and the effectiveness of feature extraction sharing mechanism.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Image Super Resolution

Image super resolution (ISR) is a process of recovering high-resolution (HR) images from its low-resolution (LR) observations. Image super resolution is also referred to by other names such as image scaling, interpolation, up-sampling, zooming and enlargement[1]. Traditional single image super resolution (SISR) restores a high-resolution image solely from its low-resolution sample. For reference-based super resolution (RefSR)[24], it utilizes the rich textures from the high-resolution references to compensate for the lost details in low-resolution images, in order to produce more realistic textures. The goal for image super resolution is to produce more subtle and natural looking textures based on the available information which can be extracted from low-resolution images.

High-resolution images provide improved reconstructed details of the scenes and constituent objects, which are critical for many devices such as large computer displays, HD television sets, and hand-held devices (mobile phones, tablets, cameras etc.). Furthermore, super-resolution has important applications in many other domains e.g. object detection in scenes (particularly small objects), face recognition in surveillance videos, medical imaging, improving interpretation of images in remote sensing, astronomical images, and forensics.[1]

1.2 Challenging

Image super resolution is still a challenging and open research problem mainly due to its ill-posed nature that the relation between high-resolution and low-resolution observation is not one-to-one correspondence. A low-resolution image can be mapped to infinite number of high-resolution images. There is no absolute result for up-scaling an image.

Because of this ill-posed nature, most existing methods suffer from blurry results at relatively large up-scaling factors, especially when fine textures are needed to be recovered. These fine textures present in the original high-resolution images are lost in low-resolution images.

1.3 Prior Works

Image super resolution is a classical task which has been studied for decades. Traditional approaches such as bicubic interpolation, sparse-coding[18] and local linear regression[19] are based on sampling theory but could only produce blurry outputs which lack detailed and realistic features.

Image super resolution has received a boost in performance with deep-learning based methods in terms of either Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) or visual quality compared to traditional non-deep-learning based methods. The introduction of convolutional neural networks (CNN) has greatly advanced the state-of-the-art

super resolution methods. A milestone work that introduced CNN into SISR was proposed by Dong et al.[17], where he trained a convolutional neural network to minimize the MSE between SR images output by the network and the original HR images. The new method achieved state-of-the-art performance compared to its non-deep-learning based counterparts. Ever since then, plenty of data-driven deep learning network architectures have been studied for image super resolution.

1.4 Motivation

To train SISR or RefSR networks, we need a great number of LR-HR image pairs as training data for generalization. Existing SISR and RefSR methods assume that ground truth is available or down-scaling is allowed to get corresponding image pairs for supervised training on the degraded scale. Typically, bicubic interpolation down-scaled input images are utilized as correspondence low-resolution images. Few research works have been done for methods that are trained in an unsupervised way. The assumption that super resolution models trained on the degraded scale are equal to that trained on the original scale is unproven. In this way, there is a hidden prior that the relation between the ground truth and low-resolution images is established manually, which may not be practical in real-world scenarios.

While these externally supervised methods perform extremely well on data satisfying the conditions they were trained on, their performance deteriorates significantly once these conditions are not satisfied anymore.[2] Besides, down-scaling will cause information lost to the original image. The training using the down-scaled images to restore the lost information of the original image is not exactly the same as generating new information directly on the original image. Although Internal Image Statistics[9], in other words, the recurrence of small pieces of information across scales of a single image, is shown to be a very strong property of natural images, there are cases where we need scale-variant information and we need to generate new information based on stationary scale.

A typical example is satellite images. Unlike natural images, where the distance between the camera and the object is flexible, for satellite images, the distance between a certain satellite and the ground is fixed. Every pixel in a satellite image refers to a small area on the ground. We would like to see the up-scaling result overfitted directly on that scale rather than on its degraded scale. It is not to say that satellite images do not comply with the Internal Image Statistics, what we would like to see is whether utilizing scale-variant information together with the cross-scale features could produce better outputs.

In cases where LR-HR image pairs are not available, unsupervised image super resolution techniques tend to be the way out. However, for unsupervised image super resolution, the biggest problem is how to create a cross-scale information transformation process to generate additional information in super-resolved images, especially high frequency information. This process is straightforward for supervised training, because there are ground truth target images to design loss functions. The super-resolved output can compare with the ground truth directly to guide the network training. If ground truth is not available, generating high-quality realistic scale-invariant features is a difficult problem.

Existing unsupervised learning methods all involve supervised training to some extent: ZSSR[7] uses patches extracted from the original image as high-resolution samples and down-

scales these patches to obtain low-resolution samples; SinGAN[5] down-scales the original image to obtain image pyramid, image with larger scale in the pyramid is regarded as the ground truth for the consecutive image with smaller scale; Unsupervised Super Resolution Cycle-in-Cycle GAN[2] uses external LR-HR image pairs through down-scaling the image meant for adversarial training.

The goal of our thesis work is to devise a pipeline for reference-based self-supervised image super resolution task where ground truth is not available and training on the degraded scale is not allowed. Existing SISR methods fail to meet this requirement. If we down-scale the original images to get low-resolution samples while the original images serve as ground truth, then the training on low-resolution samples is on the wrong scale. For current unsupervised methods, they also fail to meet this requirement since they are trained in a supervised way as SISR methods, for example, ZSSR uses patches extracted from the original image as high-resolution samples and down-scales these patches to obtain low-resolution samples where training is conducted.

1.5 Contribution

Our thesis work mainly discusses reference-based self-supervised image super resolution. Unsupervised image super resolution and supervised learning on the correct scale are studied. The main contributions of our work are:

- We propose an end-to-end pipeline for reference-based self-supervised image super resolution problem. Low-resolution images and references are input during training, while testing could be done for single image scenario.
- We propose optical-flow-based cycle-consistent image super resolution pipeline where CycleGAN[3] is adopted for style transformation between high-resolution images and super-resolved images, and optical flow estimation is adopted for multi-view information learning.
- We propose CircleSRNet where a feature extraction sharing mechanism is introduced. We further propose FlowCircleSRCycleGNet by adopting this mechanism into our pipeline. We demonstrate its superior performance compared to state-of-the-art SISR methods.

1.6 Organization

We first discuss related works on SISR, RefSR and unsupervised super resolution in chapter 2. Then, in chapter 3, our proposed pipeline is introduced, as well as its settings. In chapter 4, we test our models qualitatively and quantitatively to explore possible improvement. Chapter 5 provides a summary of conducted work, as well as future expectations.

2. Related Works

Since traditional non-deep-learning based methods such as bicubic interpolation, sparse-coding[18] and local linear regression[19] are out-performed by their deep-learning based counterparts, we only present a brief mathematical background but mainly focus on deep-learning based algorithms.

We divided deep-learning based algorithms into three categories: 1) Single Image Super Resolution (SISR), 2) Reference-Based Image Super Resolution (RefSR) and 3) Unsupervised Image Super Resolution. SISR is the most important section since it is the foundation of RefSR and unsupervised methods.

2.1 Mathematical Background

Suppose a low-resolution image is denoted by Y and the correspondence high-resolution image is X, then the degradation process is defined as:

$$Y = \phi(X, \Theta_d) \qquad (2-1)$$

Θ_d is the parameters for down-scaling. ϕ and Θ_d are unknown since the down-scaling process is hidden. Then the problem is to find a suitable reverse map which maps Y to X to recover an approximation of X.

$$\hat{X} = \phi^{-1}(Y, \Theta_u) \qquad (2-2)$$

Θ_u is the parameters for up-scaling. Since there are plenty of choices, we need to make sure that our approximation meets the requirement that it has the same content as original LR image plus it has high visual quality and fine textures. The aim is to minimize the data fidelity term with regularizers.

$$Loss(X, \Theta_d, \Theta_u) = \|\phi(X, \Theta_d) - Y\| + \lambda\varphi(X, \Theta_u) \qquad (2-3)$$

$\|\phi(X, \Theta_d) - Y\|$ is the data fidelity term, $\varphi(X, \Theta_u)$ is the regularizer and λ is the weight. The down-scaling process can be predefined as interpolation in most cases. Our main goal is to train deep learning networks to learn the prior for up-scaling thus we could reconstruct high-quality super-resolved images.

2.2 Single Image Super Resolution (SISR)

Single image super resolution (SISR) generates high-resolution images from low-resolution samples without additional references. There are dozens of promising deep neural networks for single image super resolution with different structures. We could not get through all of them but can focus on some state-of-the-art representatives. A relatively complete overview of SISR methods could be shown in Figure 2.1 from the survey[1].

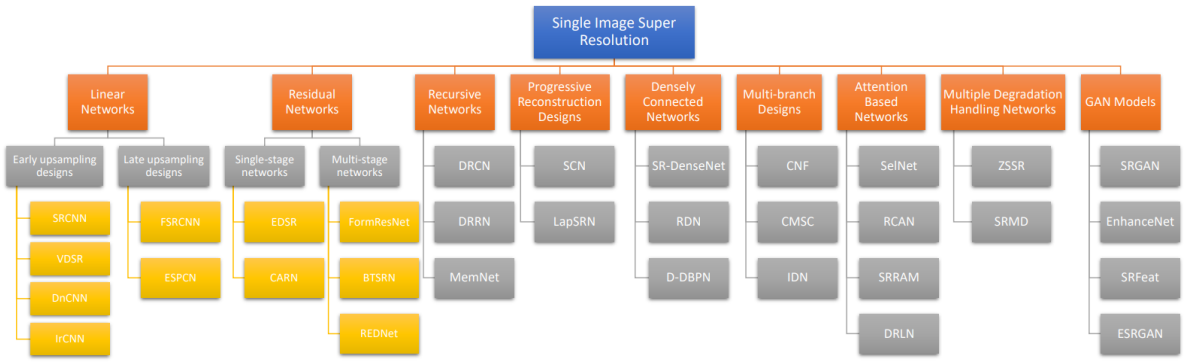


Figure 2.1: The taxonomy of the existing single-image super-resolution techniques based on the most distinguishing features[1].

2.2.1 Enhanced Deep Super Resolution (EDSR)

The enhanced deep super resolution network (EDSR)[15] is a modification of the previous work of ResNet[20] where deep residual network is adopted. Specifically, the author demonstrated dramatic improvement of performance by removing Batch Normalization Layers in residual blocks and ReLU activation. It is better to remove batch normalization layers because they get rid of range flexibility from networks by normalizing the features. This modification led to faster computation and deeper structure with relatively smaller size. Since the scale-changing layers are at the beginning and the end of the network, this structure can also be extended to multiple scales (MDSR) through shared residual blocks easily.

2.2.2 EnhanceNet

Regular image quality criteria such as PSNR do not comply with the perceptual quality of an image, causing over smoothed results in favor for a higher PSNR score. EnhanceNet[12] adopted two extra loss terms to guide the network training for a better visual quality with finer texture details. One loss is perceptual loss that measures high-level perceptual and semantic differences between images. This high-level perceptual feature can be represented as intermediate results of a pretrained network like VGG[21], which is believed to main for image feature extraction. The other is texture matching loss that measures patch-wise difference of gram matrices of deep features from a pre-trained network such as VGG. The network therefore learns to produce images that have the same local textures as the high-resolution images during training. EnhanceNet is adversarial trained for more realistic outputs.

2.2.3 Residual Channel Attention Network (RCAN)

Residual Channel Attention Network (RCAN)[22] adopts a recursive residual design where residual connections exist within each block of a global residual network. Besides, each local residual block has a channel attention mechanism to weight different channels for channel selection. The residual-in-residual structure allows multiple pathways for information flow in

the network and the channel attention mechanism guides the network to focus on more important feature maps.

2.2.4 Densely Residual Laplacian Attention Network (DRLN)

Densely Residual Laplacian Attention Network (DRLN)[23] adopts modular architecture with Laplacian attention. The improvement of DRLN[23] compared to RCAN[22] can be attributed to the cascading structure and Laplacian attention. DRLN[23] reduces the number of convolutional layers leading to a faster training.

2.3 Reference-Based Image Super Resolution (RefSR)

In contrast to SISR, RefSR methods introduce additional high-resolution reference images to assist the super resolution process. Typically, these reference images share similar content or textures with the low-resolution images like adjacent frames in a video or images from different view angle. Essentially, how to transfer the high-frequency details from the reference image to the LR image is the key to the success of RefSR. There are two main issues for RefSR: a) Building correspondence between the target HR image and the reference image, b) synthesizing high resolution output from LR image[16].

2.3.1 CrossNet

CrossNet[16] is an end-to-end convolutional neural network for RefSR. It adopts the idea of warping to build the correspondence between the target HR image and reference image and uses Encoder-Decoder structure for the synthesis of high-resolution output. The Warping result is concatenated into the decoder to assist the SR process.

Specifically, it uses optical flow estimation to generate the cross-scale correspondence between the super-resolved LR image and its corresponding high-resolution reference image. Meanwhile, it uses encoders to extract image feature maps at multiple scales of super-resolved LR image and its reference. Then, there is a decoder which performs multi-scale feature fusion and synthesis using the U-Net[25] structure with encoders.

They demonstrated the superior performance of CrossNet compared to previous SISR methods with the help of information in reference images.

2.3.2 Image Super-Resolution by Neural Texture Transfer (SRNTT)

In research work such as CrossNet, there is a strong assumption that the references have to be well-aligned to the LR image. The performance of the model highly depends on how well the references could be aligned. Besides, optical flow estimation is limited in matching long-distance correspondences, thus it is incapable of handling significantly misaligned references. If the reference image fails to align with the LR image, the performance will significantly degrade and could even become worse than SISR methods. To address these problems, it

proposed Neural textural Transfer[26] which conducts local texture matching in the feature space and transfer matched textures to the final output.

2.4 Unsupervised Image Super Resolution

When paired LR-HR image data is unavailable, unsupervised learning is needed. GAN[11] is widely used to solve the unsupervised learning problems, which typically includes a generator and a discriminator. The generator is trained to generate fake images to fool the discriminator while the discriminator tries to distinguish the generated results from the real data. There are few research works related to unsupervised image super resolution since down-scaling is widely accepted to get LR-HR image pairs.

2.4.1 "Zero-Shot" Super-Resolution using Deep Internal Learning (ZSSR)

Based on Internal Image Statistics[4][9] that similar image patches tend to repeat inside a single image both within the same scale as well as across different image scales, the internal entropy of patches inside a single image is much smaller than the external entropy of patches in a general collection of natural images. Thus, to super resolve a LR image, the feature information inside itself is more concentrated and relevant than features in a broad image collection. Therefore, training an image-specific CNN on patches extracted solely from input image at test time is reasonable[7]. Specifically, given a certain test image, ZSSR[7] trains a small CNN on LR-HR patch pairs extracted from this image, and then applies this CNN to the original image itself to produce the super-resolved output.

2.4.2 SinGAN: Learning a Generative Model from a Single Natural Image

SinGAN[5] is an unconditional generative model that can capture the internal statistics of a single image based on a multi-scale architecture done in a coarse-to-fine fashion. It consists of a pyramid of generators which are trained against an image pyramid, where this image pyramid is obtained by down-scaling the original image by a fixed factor r. Each generator is responsible for generating a realistic image sample with regard to the corresponding image in the image pyramid through adversarial training. The generation of an image sample starts at the coarsest scale and sequentially passes through all generators up to the finest scale, with noise injected at every scale. Since the generation process is done in a coarse-to-fine fashion, SinGAN can be applied to super resolution by up-sampling the finest scale to obtain the high-resolution result.

2.4.3 Unsupervised Image Super-Resolution using Cycle-in-Cycle Generative Adversarial Networks

Assume that LR-HR image pairs and the down-sampling process are not available, it is difficult to perform super resolution if the input images suffer from different kinds of degradation.

Inspired by the image-to-image translation applications, their proposed model adopts two CycleGANs[3]. The first CycleGAN structure maps the input noisy LR images to the clean and bicubic down-sampled LR space. Then, a well-trained super resolution network is used to up-scale the cleaning input image to the desired scale. Finally, the second CycleGAN structure covers the first one to fine-tune the result high resolution output.

3. Proposed Model

We propose optical-flow-based cycle-consistent image super resolution pipeline for image super resolution problem when ground truth is not available and training on the degraded scale is not allowed. The same structure can be applied to both unsupervised learning and supervised learning.

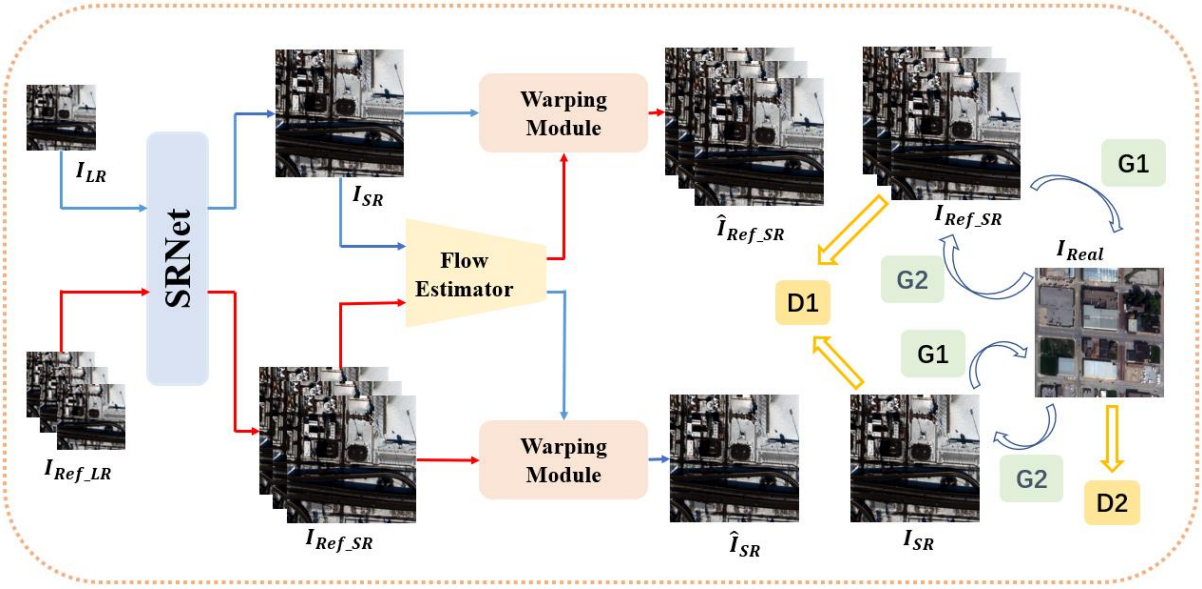


Figure 3.1: Pipeline Overview

3.1 Pipeline Overview

As we focus on self-supervised image super resolution, our proposed model utilizes additional viewpoints to help super resolve the low-resolution input data. There are multiple images sharing similar content which can be used to build our reference-based structure. The pipeline overview is shown in Figure 3.1. There are four modules:

a) Super Resolution Network (SRNet)
b) Optical Flow Estimator
c) Warping Module
d) Cycle-Consistent Generative Adversarial Network (CycleGAN)[10]

The Super Resolution Network super resolve input low-resolution images to corresponding high-resolution images. The Optical Flow Estimator aims to build correspondence between central image and its references, where multi-view information could be learned to guide the training of super resolution network. The Warping Module could be used to reconstruct images from its references and optical flow estimation. The CycleGAN is adopted to transfer features from real high-resolution images to super-resolved images. Initially, low-resolution central image and its low-resolution references are super resolved to obtain corresponding super-resolved central image and its super-resolved reference images, unlike typical reference-based super resolution where references are high-resolution images.

Meanwhile, bidirectional optical flow estimation between the super-resolved central image and its super-resolved references is conducted, which could be used by the warping module to build reconstruction loss. Then, CycleGAN builds an image-to-image translation between real high-resolution images and super-resolved images to polish the performance of super resolution network, where G1 transforms image from super-resolved domain to high-resolution domain, G2 transforms image from high-resolution domain to super-resolved domain, D1 is the discriminator for super-resolved domain and D2 is the discriminator for high-resolution domain.

3.2 Super Resolution Module

This module of the proposed pipeline aims in super-resolving the low-resolution images to the target scale. Existing state-of-the-art SISR methods could be applied for this module, such as EDSR[15], RCAN[22] or DRLN[23]. Typical SISR methods are trained in a supervised way where high-resolution ground truth is available. Assume there is ground truth I_{GT}, L1 loss $L_{SR} = |I_{GT} - I_{SR}|$ is widely adopted between super-resolved image I_{SR} and ground truth for a better pixel reconstruction effect.

The structure of EDSR[15] is adopted and modified for our work due to its superior performance with faster training and fewer parameters. EDSR[15] is a modification of the previous work ResNet by removing Batch Normalization Layers and ReLU activation in residual blocks. The network structure of EDSR[15] is shown in Figure 3.2.

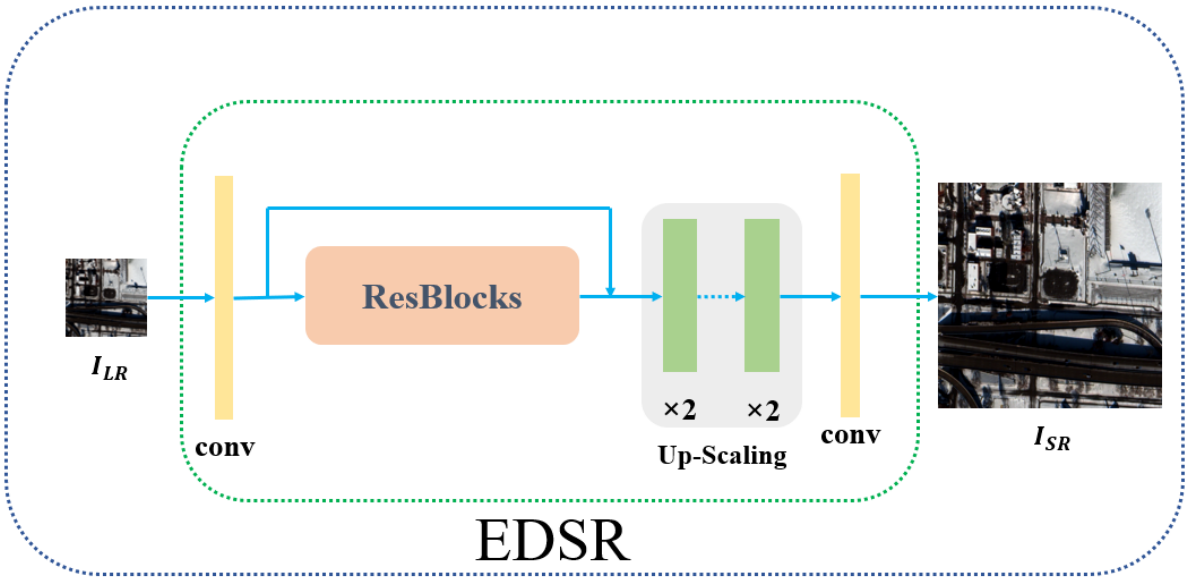


Figure 3.2: EDSR Network Structure

The performance of our pipeline highly depends on the performance of super resolution network. This module is the core of the whole pipeline and all the other modules aim to guide the training of super resolution network in an end-to-end way.

3.3 Optical Flow Estimation

This module of the proposed pipeline aims for building correspondence between the reference images and original central image. If training data present good alignment, optical flow estimation could be applied to fulfil this requirement. Bidirectional optical flow estimation strategy is adopted for better performance. With optical flow estimation, multi-view information could be utilized to guide network training. The resulting optical flow estimation could be used to synthesize the target image.

Given the central image I_{SR} and references I_{Ref_SR}, our bidirectional optical flow estimates two flows:

a) $F_{Cen} \rightarrow F_{Ref}$ optical flow from central image to reference images

b) $F_{Ref} \rightarrow F_{Cen}$ optical flow from reference images to central image

Existing optical flow network could be applied for this module, such as FlowNet[27] or PWC-Net[28]. Based on state-of-the-art flow estimation performance, PWC-Net is adopted and modified in our work. The generation of flow estimation in one layer of PWC-Net can be seen in Figure 3.3.

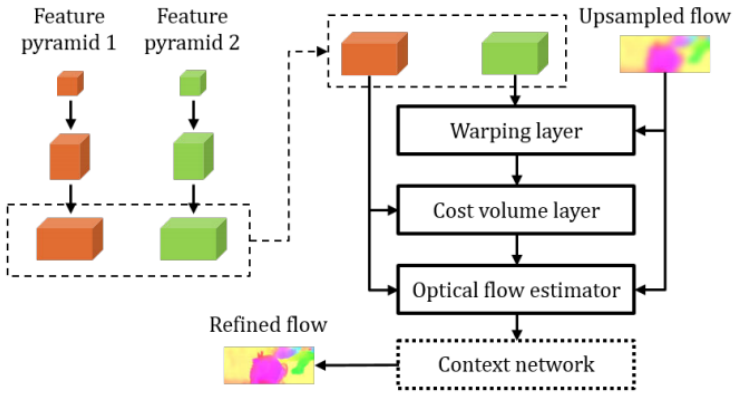


Figure 3.3: PWC-Net Optical Flow Estimation[28]

Flows in PWC-Net are generated in a coarse-to-fine manner. A certain flow is generated from correspondence feature maps of central image and its references in the feature pyramid, as well as up-sampled flow from previous step. The output of PWC-Net is the collection of these flows, while the scale of the finest flow is 4 times smaller than the input images.

3.4 Warping

The key idea of reference-based super resolution lies in synthesizing super-resolved output from the correspondence between central image and reference images. Warping is adopted in our pipeline to synthesize and reconstruct images using references and optical flow estimation.

Given central image I_{SR} and references I_{Ref_SR}, as well as optical flow estimation $F_{Cen} \rightarrow F_{Ref}$ and $F_{Ref} \rightarrow F_{Cen}$ in two directions. Synthesized images $\hat{I}_{SR}$ and $\hat{I}_{Ref\_SR}$ could be obtained by:

$$\hat{I}_{SR} = Wrapper\left(I_{Ref\_SR}, F_{Cen} \rightarrow F_{Ref}\right) \qquad (3-1)$$

$$\hat{I}_{Ref\_SR} = Wrapper\left(I_{SR}, F_{Ref} \rightarrow F_{Cen}\right) \quad (3-2)$$

In ideal situation, synthesized image shall be the same as original image. In real scenario, we could use mask to exclude places that could not be synthesized for cases like occlusion. Wrong synthesized pixel values hinder the performance of optical flow estimation. Mask function is used to generate a mask by forward warping the optical flow estimation. The mask should only keep positions where high-quality reconstruction is available. For central image I_{SR}, its mask could be obtained by $M_{Cen} = MF(F_{Ref} \rightarrow F_{Cen})$ where MF refers to the mask function. Similarly, mask for references I_{Ref_SR} can be generated by $M_{Ref} = MF(F_{Cen} \rightarrow F_{Ref})$. Synthesized images with mask could be obtained by:

$$\hat{I}_{SR} = Wrapper\left(I_{Ref\_SR}, F_{Cen} \rightarrow F_{Ref}\right) \cdot M_{Cen} \quad (3-3)$$

$$\hat{I}_{Ref\_SR} = Wrapper\left(I_{SR}, F_{Ref} \rightarrow F_{Cen}\right) \cdot M_{Ref} \quad (3-4)$$

Mask has same shape as its corresponding image. If reconstruction is available for a certain pixel position, mask value is 1, otherwise it is 0.

3.5 Cycle-Consistent Generative Adversarial Network

For unsupervised scenario, despite guided by the optical flow estimation, without ground truth, the output from the super resolution network still could not produce satisfactory high-quality result. The biggest problem is still how to restore the features in high-resolution scale. Optical flow estimation adds additional information from reference images. It builds a correspondence between super-resolved central image and its super-resolved reference images. However, this correspondence is based on the same scale. For typical reference-based super resolution, there are high-resolution reference images, thus building cross-scale information transformation is straightforward.

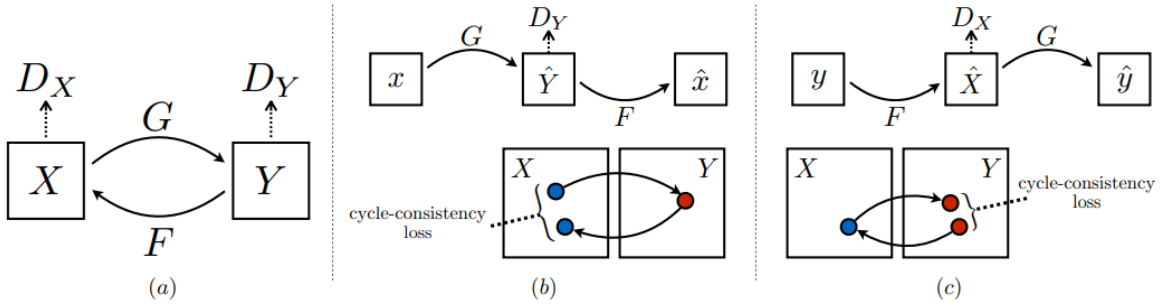


Figure 3.4: CycleGAN Structure[3]

Inspired by the research work of image-to-image translation, CycleGAN is adopted for feature transformation between super-resolved images and real high-resolution images. For CycleGAN structure, there are two mapping functions $G : X \rightarrow Y$, $F : Y \rightarrow X$ and associated adversarial discriminators D_X and D_Y as shown in Figure 3.4 (a). Two cycle-consistency losses are defined for image translation between space X and space Y as shown in Figure 3.4

(b) and (c). An image in a certain space transformed to another space and then transformed back should be the same as before transformation.

Typically, CycleGAN is used for image style transformation when space *X* and space *Y* have different style patterns. The generators learn pattern maps between different spaces. In our scenario, we would like the generators to learn from feature pattern difference between manually super-resolved images and real high-resolution images. It is not quite the same since what we need is not style difference but high visual-quality super-resolved output. It is hard to define so-called-style of super-resolved images from real high-resolution images because normally they have similar style and content but different visual quality. Thus, simply adopting CycleGAN structure could not restore features in high-resolution scale[2]. The performance of CycleGAN highly depends on the performance of super resolution network. The training of CycleGAN should comply with the training of super resolution network. A better super resolution network could produce super-resolved images with finer textures, and then CycleGAN could learn finer differences in feature patterns between the super-resolved images and real high-resolution images. Since the style difference between super-resolved images and real high-resolution images is not obvious, we should keep the content-consistency in two spaces.

3.6 Loss

As an end-to-end network structure, our model can be directly trained to synthesize super-resolved outputs given proper loss functions. Nine losses are introduced in this section:

- Self-Consistency Loss
- Back-Propagation Loss
- Perceptual Loss
- Texture-Matching Loss
- Flow-Synthetic Loss
- Adversarial Loss
- Cycle-Consistency Loss
- Identity Loss
- Frequency Loss

Among these losses, self-consistency loss, perceptual loss, texture-matching loss and frequency loss are meant for supervised training when high-resolution and low-resolution image pairs are available. Cycle-consistency loss and identity loss are included if CycleGAN structure is adopted.

3.6.1 Self-Consistency Loss

Self-consistency loss is the main loss for super-resolution. The super-resolved output shall keep the same content as the low-resolution input image. For supervised learning when low-resolution and high-resolution image pairs are available, self-consistency loss could also ensure cross-scale feature learning. For state-of-the-art SISR methods, trained with only self-consistency loss is adequate to generate relatively high-quality super-resolved output. Self-consistency loss is in pixel level across whole image.

Given super-resolved central image I_{SR}, references I_{Ref_SR}, and associated ground truth I_{GT} and I_{Ref_GT}, self-consistency loss can be denoted by:

$$L_{SC}^{cen} = \|I_{SR} - I_{GT}\|_1 \tag{3-5}$$

$$L_{SC}^{ref} = \left\|I_{Ref\_SR} - I_{Ref\_GT}\right\|_1 \tag{3-6}$$

If we adopt down-scaling process to get correspondence image pairs, then I_{GT} refers to the original input image and I_{SR} is the super-resolved image from the degraded input.

3.6.2 Back-Propagation Loss

If ground truth is not available, self-consistency loss between the ground truth and output could not be established. To ensure data fidelity, back-propagation loss is adopted. The down-scaled output image should comply with the original input.

Given input I_{LR}, its references I_{Ref_LR}, and down-scaled super-resolved central image $\hat{I}_{LR}$, references $\hat{I}_{Ref\_LR}$, back-propagation loss could be given by:

$$L_{BP}^{cen} = \left\|I_{LR} - \hat{I}_{LR}\right\|_1 \tag{3-7}$$

$$L_{BP}^{ref} = \left\| I_{Ref\_LR} - \hat{I}_{Ref\_LR} \right\|_1 \quad (3-8)$$

3.6.3 Perceptual Loss

Empirically, ensuring pixel level consistency is not enough for the output to have state-of-the-art high-quality visual presentation. The output tends to suffer from blurry although it may have high PSNR. Instead of encouraging pixels of output image to exactly match pixels of target image, we prefer that output image has similar feature representations as target. We adopt perceptual loss[14] that measures high-level perceptual and semantic differences between images. A pre-trained deep loss network φ is utilized to represent high-level feature maps of images. In our work, φ is a pretrained VGG[21] network.

The perceptual loss is then defined as Euclidean distance between feature representations:

$$L_P^{cen} = \|\varphi(I_{SR}) - \varphi(I_{GT})\|_1 \quad (3-9)$$

$$L_P^{ref} = \left\| \varphi\left(I_{Ref\_SR}\right) - \varphi\left(I_{Ref\_GT}\right) \right\|_1 \quad (3-10)$$

3.6.4 Texture-Matching Loss

According to recent works on image style transformation, convolutional neural networks can be used to create high-quality textures. Given a target texture image, the output image is generated iteratively by matching statistics extracted from a pre-trained network to the target texture.

Here we adopt Gram matrix $G(X) = XX^T$ as the feature statistics to match. To gain feature maps, a pre-trained VGG[21] network is used to calculate the feature maps of given images. These feature maps are transformed through Gram matrix to be the texture presentation. The texture loss[12] is then defined as the Euclidean distance between the Gram matrix of feature maps:

$$L_T^{cen} = \left\| G\left(\varphi(I_{SR})\right) - G\left(\varphi(I_{GT})\right) \right\|_2 \quad (3-11)$$

$$L_T^{ref} = \left\| G\left(\varphi\left(I_{Ref\_SR}\right)\right) - G\left(\varphi\left(I_{Ref\_GT}\right)\right) \right\|_2 \quad (3-12)$$

Instead of apply this texture matching formulation globally for high resolution textures, we calculate texture loss patch-wise during training to enforce locally similar textures between the output and the target. Our network is trained to approximate similar local features from the corresponding high-resolution images.

3.6.5 Flow-Synthetic Loss

The additional information of reference images is utilized through optical flow estimation. We could synthesize and reconstruct images using references and corresponding flow estimation through warping. The reconstructed image should approximate the original image as close as possible. We could minimize the photometric loss[29] by associating the pixel intensities

between different views. Mask is introduced to exclude regions where correspondence is failed to establish.

Given super-resolved central image I_{SR}, references I_{Ref_SR}, and associated synthetic images using optical flow estimation $\hat{I}_{SR}$ and $\hat{I}_{Ref\_SR}$, the flow-synthetic loss is given by:

$$L_{FS}^{cen} = \left\| I_{SR} - \hat{I}_{SR} \right\|_1 \cdot M_{Cen} \quad (3-13)$$

$$L_{FS}^{ref} = \left\| I_{Ref\_SR} - \hat{I}_{Ref\_SR} \right\|_1 \cdot M_{Ref} \quad (3-14)$$

3.6.6 Adversarial Loss

Adversarial learning strategy[11] is used during the model training process to generate more realistic output images. For adversarial training, typically there is a generator G and a discriminator D. The generator tries to generate satisfactory fake data while the discriminator tries to distinguish this fake data from real data. Typical Adversarial loss is denoted by:

$$L_D = -\log\big(D(x)\big) - \log\Big(1 - D\big(G(z)\big)\Big) \quad (3-15)$$

x is a sample from real data distribution, z is random noise as the input into the generator. For conditional GAN, z could be replaced by a specific given image, like a low-resolution image in our case. The super resolution network could be regarded as a generator.

3.6.7 Cycle-Consistency Loss

In our work, we adopt CycleGAN[3] structure apart from traditional GAN structure. According to the pipeline overview in Figure 3.1 as well as CycleGAN structure in Figure 3.4. There are two generators G1 and G2, G1 translates images from super-resolved domain into high-resolution domain while G2 translates images in an opposite way. Discriminator D1 is introduced to distinguish whether a given image is from super-resolved domain or not, on the other hand, discriminator D2 is meant for classifying images of high-resolution domain. Two cycle-consistency losses are defined for bidirectional translation. A super-resolved image translated by G1 into high-resolution domain and then translated back by G2 into original domain should keep the same content, and vice versa. The cycle-consistency loss[3] could be given by:

$$L_{CC}^{G1G2} = \left\| I_{SR} - G2\big(G1(I_{SR})\big) \right\|_1 + \left\| I_{Ref\_SR} - G2\left(G1\big(I_{Ref\_SR}\big)\right) \right\|_1 \quad (3-16)$$

$$L_{CC}^{G2G1} = \left\| I_{Real} - G1\big(G2(I_{Real})\big) \right\|_1 \quad (3-17)$$

3.6.8 Identity Loss

Although adversarial training could learn maps which could produce outputs identically distributed as target domains, the introduction of CycleGAN is not to replace the function of super resolution network. Generating high-quality super-resolved features is different from style transformation where the diverse of the patterns is obvious. In fact, the performance of CycleGAN highly depends on the performance of super resolution network. We adopt identity loss to ensure the content preservation before and after style transformation. In other applications, identity loss is also adopted to preserve color composition between input and output images when working on painting generation. The identity loss[2] could be denoted by:

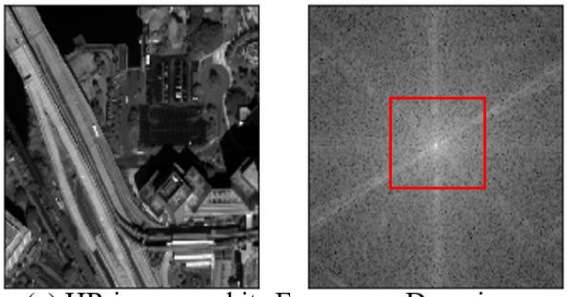
$$L_{ID}^{cen} = \|I_{SR} - G1(I_{SR})\|_1 \quad (3-18)$$

$$L_{ID}^{ref} = \left\|I_{Ref\_SR} - G1\left(I_{Ref\_SR}\right)\right\|_1 \quad (3-19)$$

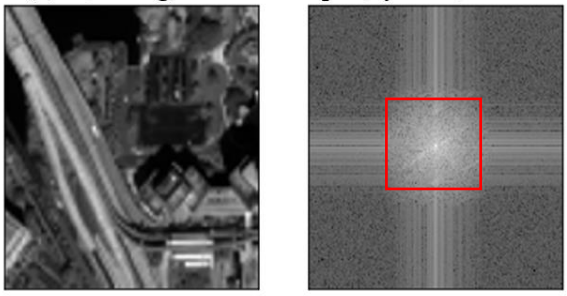
3.6.9 Frequency Loss

The Fourier Transform is an important image processing tool which is used to decompose an image into frequency domain from equivalent spatial domain. In the frequency domain image, each point represents a particular frequency contained in the spatial domain image.

Generating sufficient high frequency information in the high-resolution space is the key challenge for super resolution problem, as can be seen from Figure 3.5. For frequency domain image, points represent lower frequency information lie closer to the center of the image. Signals in the red rectangle area of Figure 3.5 refers to low frequency information. Brighter points represent stronger information. According to image frequency analysis, traditional bicubic interpolation could not restore high frequency information correctly. The pattern in frequency domain for bicubic interpolation is different compared to the ground truth.



(a) HR image and its Frequency Domain



(b) Bicubic Interpolation and its Frequency Domain

Figure 3.5: Image Frequency Analysis

In order for our network to output image which could have same information in the frequency domain as ground truth, we also add a frequency loss:

$$L_{Fre}^{cen} = \|F(I_{SR}) - F(I_{GT})\|_2 \quad (3-20)$$

$$L_{Fre}^{ref} = \left\|F\left(I_{Ref\_SR}\right) - F\left(I_{Ref\_GT}\right)\right\|_2 \quad (3-21)$$

4. Experiments

In this chapter, we introduce our dataset in section 4.1 and implementation details in section 4.2. We define our metrics in section 4.3. In section 4.4, three state-of-the-art SISR methods (EDSR[15], RCAN[22] and DRLN[23]) are adopted as baseline, as well as two additional methods (NRGSR and GESR) defined in section 4.4. We show optical flow estimation performance improvement achieved by adopting Up-Sampled PWC-Net in section 4.5. We test two unsupervised methods (UnFlowSRNet and UnFlowSRCycleGNet) in section 4.6 and two supervised methods (FlowCircleSRCycleGNet and UnFlowSRCycleGNet with external image pairs) in section 4.7. Both unsupervised and supervised methods utilize the same pipeline proposed in Chapter 3 as seen in Figure 3.1 except otherwise stated. In section 4.8 and 4.9, we present qualitative and quantitative performance overview.

4.1 Datasets

The training dataset for this work consists of 284 satellite images with size 1920×1920 categorized into 5 groups, where each group is a set of images taken on the same spot in different time of the year. One additional group of 35 1920×1920 satellite images is set aside for testing.

Since images in the same group have same location, they present good alignment. However, the changing of time can change the appearance, for example, snow can cover the ground in winter, the color of trees turns yellow in autumn and green in spring. The changing of time could also change textures, for example, cars on the roads or in the parking lots are different at different time. Such kind of changing is different from the assumption of optical flow, where we assume brightness constancy constraint. Empirically, satellite images have rich fine textures compared to natural images, making satellite image super resolution an even more challenging problem.

4.2 Implementation Details

We implement our network in PyTorch. Networks are trained 400 epochs with initial learning rate 10^{-4} and decayed to its 0.618 every 100 epochs. Random crop is conducted from the original satellite images to obtain proper size of patches to train models. We adopt batch size 4, which means 4 reference patches for a single central patch are extracted. The central patch is copied 4 times to comply with the size. For hyperparameters of losses, the weight for self-consistency loss and back-propagation loss is 1, the weight for perceptual loss is 0.04, the weight for texture-matching loss is 10^{-3}, the weight for flow-synthetic loss is 0.1, the weight for adversarial loss is 10^{-4}, the weight for cycle-consistency loss is 0.01, the weight for identity loss is 10^{-4} and the weight for frequency loss is 10^{-3}.

4.2.1 U-Net Generator or Resnet Generator for CycleGAN

We adopt U-Net structure rather than residual structure for our generator of CycleGAN. During testing, we found U-Net structured generator could have better convergence and slightly better performance. Besides, the main structure of our super resolution network is residual blocks, attaching more residual structures at the end of our super resolution network is more like repetition. After all, the super resolution network itself can be seen as generators.

4.2.2 With or Without Batch-Norm for Discriminators

The superior performance of EDSR[15] highly depends on the removal of batch normalization layers in residual blocks. Inspired by this, we would like to see possible improvement by removing batch normalization layers in our discriminators, since the performance of discriminators for CycleGAN determines how much improvement could be made to our super-resolved outputs compared with real high-resolution images. During testing, Batch-Norm based discriminators are more stable for our network, so we keep batch normalization layers.

4.2.3 PatchGAN[33] or WGAN-GP[32]

We turn to WGAN-GP for a more stable training process, given that generative adversarial networks are typically hard to train. The pattern difference between high-quality super-resolved images and real high-resolution images is particularly small as we improve the performance of our super resolution network. A stronger generator is needed to learn finer texture difference, as well as stronger discriminators to detect negligible difference between these two image sets. During training, 10 steps of training for discriminators are conducted between each time we update the parameters of our generators.

4.2.4 Reduced Cycle-Consistency Loss

For loss function of CycleGAN, the cycle-consistency term L_{CC}^{G2G1} is hard to control since there is no absolute definition of real high-resolution images. The I_{Real} we used in our training is patches randomly extracted from satellite images with scales as super-resolved images. Strictly speaking, these so-called real images is on the original scale rather than high-resolution space. The features inside could not fully represent features in high-resolution space. Less artificial textures could be made if we remove L_{CC}^{G2G1}.

4.3 Evaluation Metrics

The experiments are analyzed both qualitatively and quantitatively. All the experiments are conducted with up-scaling factor of 4. We test peak-signal-to-noise (PSNR) and structural similarity index (SSIM) values as for quantitative analysis, while intuitional visual quality is adopted for qualitative assessment.

The performances are tested on two scales, the original scale and the degraded scale. For experiments conducted on original scale, since there is no ground truth for super-resolved images from original scale, only qualitative assessments are present. On the other hand, for experiments conducted on degraded scale, the original input images can be seen as ground truth, thus quantitative measures PSNR and SSIM are included apart from visual quality assessment.

4.4 Baseline

We compare our network against three state-of-the-art SISR methods (EDSR[15], RCAN[22] and DRLN[23]), as well as two additional networks: a) Noise Residual Generator Super Resolution Network (NRGSR) and b) Gated Embedded Super Resolution Network (GESR). Although there are dozens of SISR methods, EDSR[15], RCAN[22] and DRLN[23] shall be good representatives for their superior performance against other SISR methods.

We train EDSR[15], RCAN[22], DRLN[23], NRGSR and GESR on our dataset with self-consistency loss, perceptual loss, texture-matching loss, adversarial loss and frequency loss to make sure they can performance well on our dataset. No references are included since they are SISR methods.

The qualitative performance is shown in section 4.8, while quantitative performance is shown in section 4.9.

4.4.1 Noise Residual Generator Super Resolution Network (NRGSR)

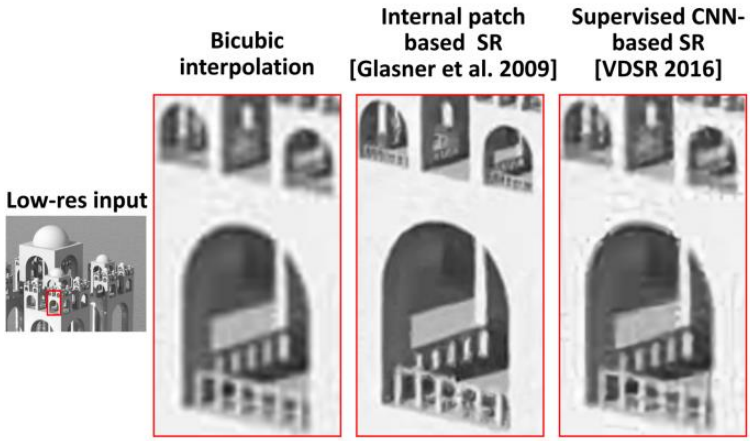


Figure 4.1: Power of Internal Image Statistics[9]

Natural image tends to have strong internal data repetition, which is the key to self-supervised learning. Empirically, Internal entropy of patches inside a single image is much smaller than the external entropy of patches in a general collection of natural images. For super resolution, learning textures and features inside the input itself is more efficient than learning similar features in a broad image collection as can be seen in Figure 4.1.

Inspired by the work of SinGAN[5] and Deep Image Prior[13], A Noise Residual Generator Super Resolution Network (NRGSR) is proposed for Single Image Super Resolution

at test stage. NRGSR is an image-specific network that only trained on a single LR-HR image pair. Based on the theory of Internal Image Statistics, features and information inside a single image is more essential and intrinsic for super resolution than features learned from a broad database.

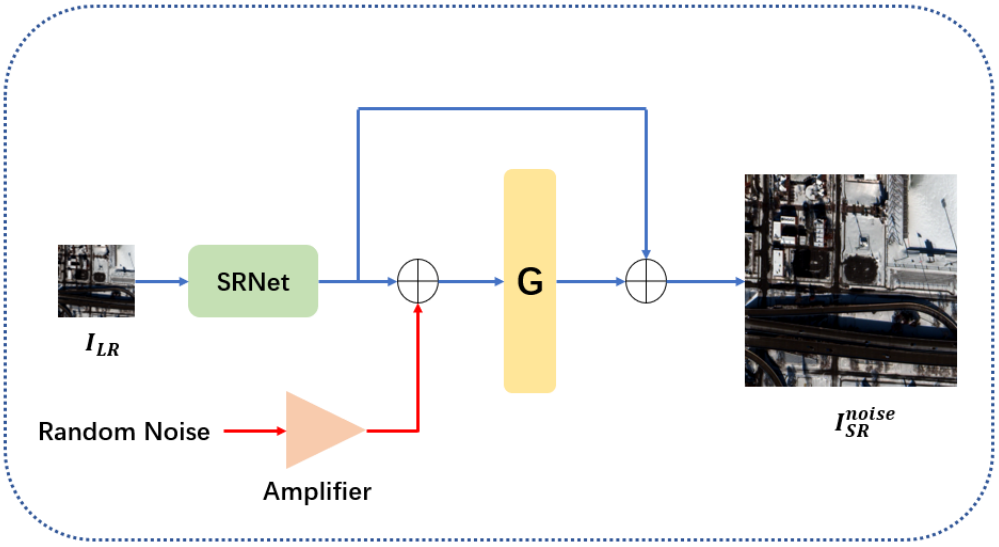


Figure 4.2: Noise Residual Generator Super Resolution Network (NRGSR)

The structure of NRGSR is shown in Figure 4.2. Baseline EDSR[15] network is adopted as the SRNet. A random noise is added on the super-resolved image to refine the super resolution process. The generator G should be a shallow network to reduce overfitting.

$$I_{SR}^{noise} = G(I_{noise} + I_{SR}) + I_{SR} \qquad (4-1)$$

We train it both on one LR-HR image pair and on the entire training dataset with the same setting. Since we use EDSR[15] as SRNet, we denote the model trained on one LR-HR image pair as NRGSR, while the model trained on entire dataset as EDSR[15]+NRG.

Although we train NRGSR on down-scaled input and there is only one image pair without further data augmentation, when applied on the original scale, it still could produce high-quality outputs as shown in Figure 4.3. SinGAN[5] tends to generate artificial textures while NRGSR highly comply with the original information.

The Noise Residual Generator structure could also improve SISR methods by sharping edges of super-resolved outputs, as can be seen in Figure 4.4.

If we remove Noise Residual Generator and train Baseline EDSR[15] directly on single image pair, the trained model failed to super resolve the original image as seen in Figure 4.5.

(a) Bicubic (b) NRGSR

(c) SinGAN[5] (d) EDSR[15]

Figure 4.3: Performance of NRGSR on Original Scale

(a) EDSR[15] + NRG (b) EDSR[15] (c) Difference

Figure 4.4: Performance Improved by Noise Residual Generator when Trained on whole Dataset

(a) EDSR[15] (b) NRGSR

Figure 4.5: Performance Improved by Noise Residual Generator on Single Image Training

4.4.2 Gated Embedded Super Resolution Network (GESR)

To increase the network's ability of learning and generating higher level features, inspired by the work of Attention U-Net[30], we introduce GESR, which is formed as a gated U-Net structure. The structure of GESR is shown in Figure 4.6.

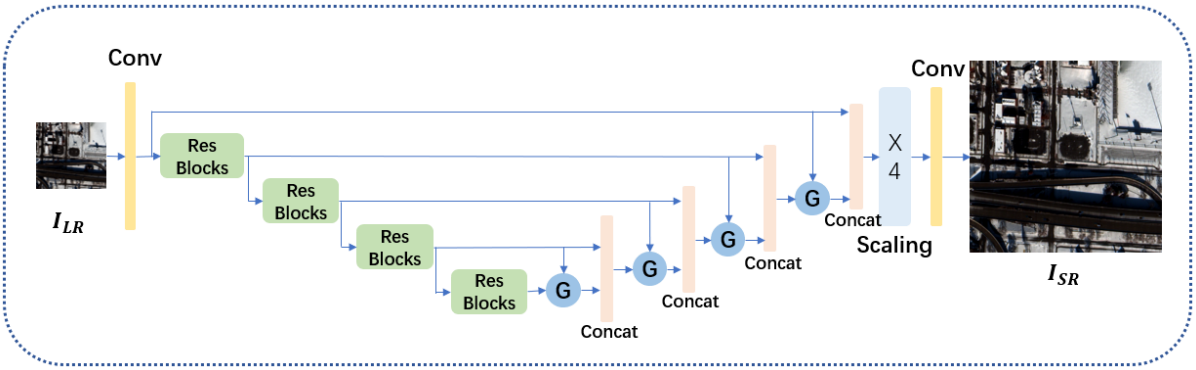


Figure 4.6: Gated Embedded Residual Super Resolution Network (GESR)

The structure of gate is shown in Figure 4.7, where signal g controls the level of signal x^l.

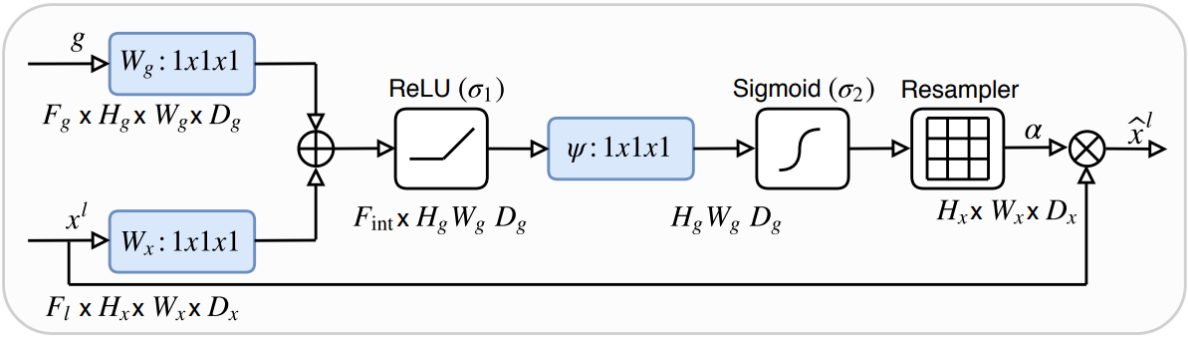


Figure 4.7: Gate Structure[30]

Deeper modules learn higher level features. The input to deeper residual blocks is the output of previous residual blocks, so it shall control the weight of how much higher-level information need to be added.

4.5 Up-Sampled Optical Flow Estimator

We adopt the widely used PWC-Net[28] as our flow estimator to build the correspondence between the central image and its references. Unlike traditional reference-based image super resolution where high-resolution reference images are available, in our situation, reference images are low-resolution images with similar content as central image. Flow estimation is built for the multi-view information under same scale rather than for cross-scale feature information transformation.

The performance of multi-view information learning highly depends on the performance of optical flow estimation. Unlike routine optical flow estimation applications where motion flow is continuous, such as consecutive frames of the video, satellite images taken in different time can have massive difference. Intensities for same voxels in two satellite images could change dramatically. To further improve the performance of PWC-Net, we replace the final $\times 4$ bilinear up-sampling process with 2 additional up-sampling module the same as previous up-sampling steps, as seen in Figure 4.8. This modification ensures the flow estimation to adapt to finer texture changes.

If central image presents good alignment with the reference image, PWC-Net optical flow estimation could synthesize high-quality image to approximate the original central image as shown in Figure 4.9. However, because of intensity changing for same voxel, synthetic image has different appearance from original image, like different color of trees. Besides, if such alignment is not available, central image synthesized may fail to approximate the original one as shown in Figure 4.10.

For Up-Sampled PWC-Net, it could measure finer texture changes and criterion for building the correspondence is way more strict than original PWC-Net. The estimation result can be seen in Figure 4.11. Mask of Up-Sampled PWC-Net could eliminate patches where alignment fails to meet our requirement. For the same voxel, it is also masked out if it has different intensity. More pixels are masked out than previous PWC-Net to produce high-quality synthetic output. We could observe a significant improvement of MSE loss for reconstruction as shown in Table 4.1.

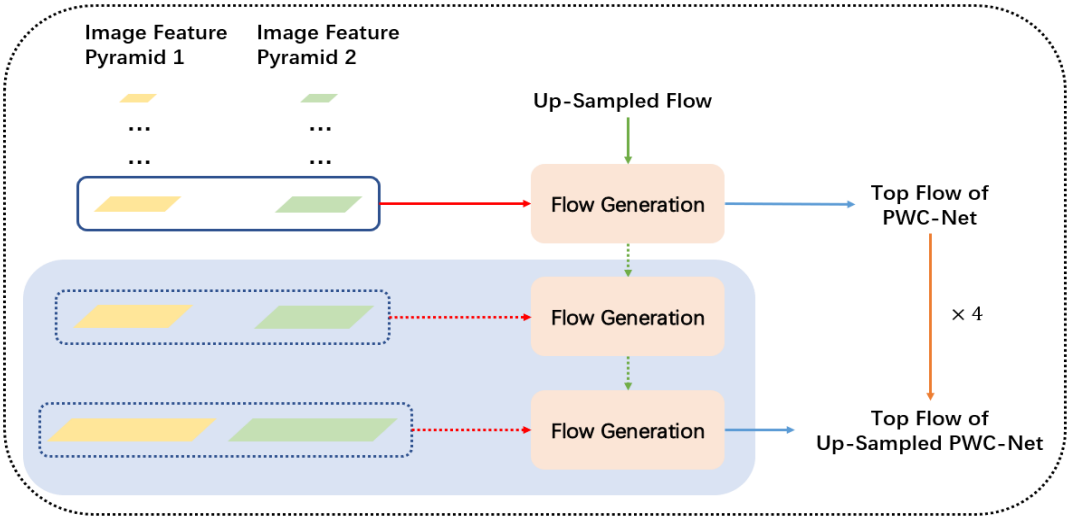
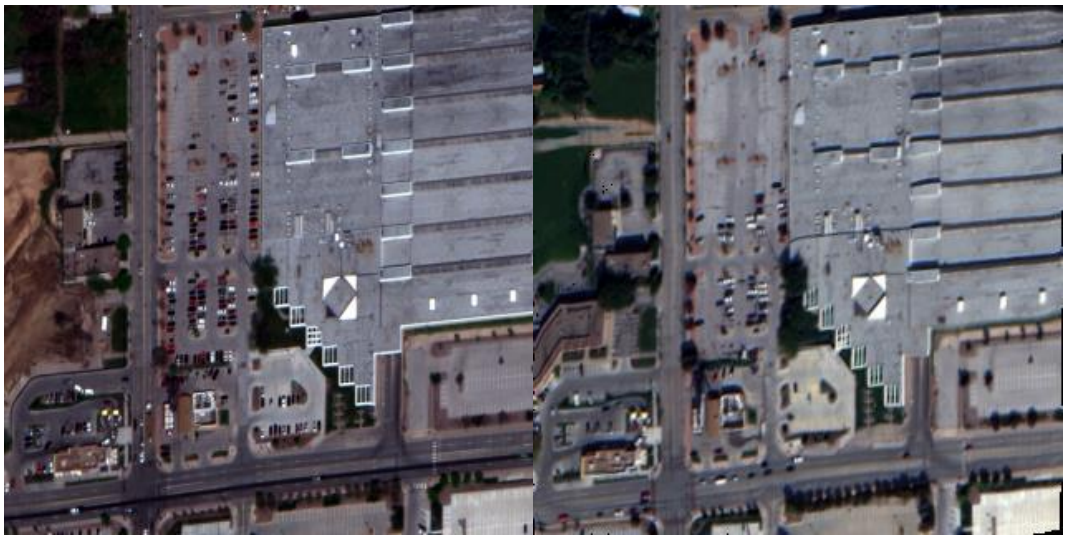


Figure 4.8: Up-Sampled PWC-Net



(a) Ground Truth (b) Synthetic Image

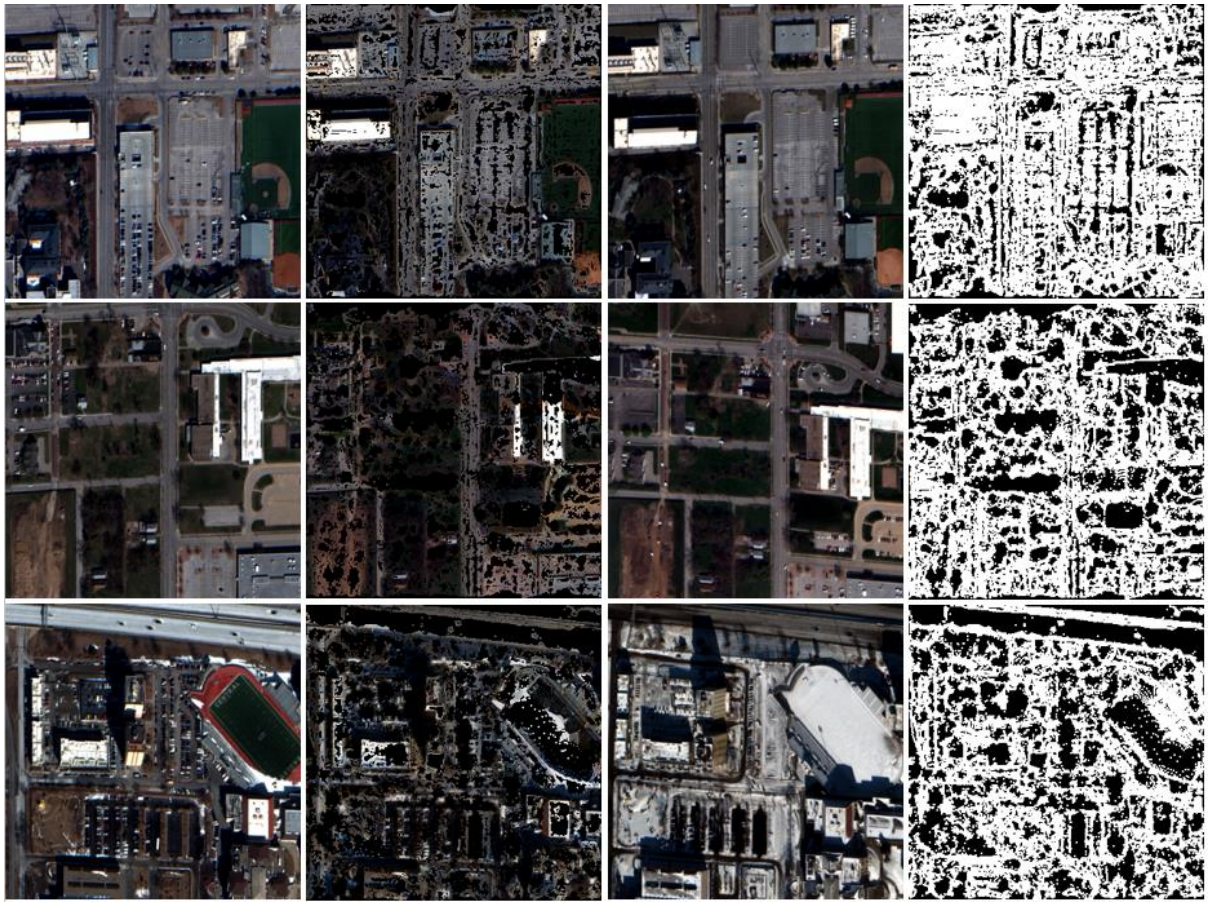
(c) Reference (d) Mask

Figure 4.9: PWC-Net Optical Flow Estimation with Good Alignment

(a) Ground Truth (b) Synthetic Image

(c) Reference (d) Mask

Figure 4.10: PWC-Net Optical Flow Estimation with Bad Alignment



(a) Ground Truth (b) Synthetic Image (c) Reference (d) Mask

Figure 4.11: Up-Sampled PWC-Net Optical Flow Estimation

Models	MSE
PWC-Net	0.0784
Up-Sampled PWC-Net	0.0188

Table 4.1: MSE of PWC-Net and Up-Sampled PWC-Net

4.6 Unsupervised Learning

Pure reference-based unsupervised learning restores features in high-resolution space from low-resolution central images and low-resolution reference images without scaling the input images. If scaling could be conducted, then we could conveniently generate LR-HR image pairs and apply relative supervised methods for image super resolution.

Due to the special natural of satellite images, where down-scaling may cause massive scale-relevant feature loss, we propose unsupervised optical-flow-based super resolution network (UnFlowSRNet) and its modification by adding CycleGAN (UnFlowSRCycleGNet) to address this problem.

For unsupervised learning on the original scale, 32×32 size patches are extracted on the central image, while 128×128 size patches randomly extracted on the same image are regarded as real high-resolution images for adversarial training. For references, patches with same size and same pixel location are extracted from random-shuffled neighboring images within same group to ensure the extracted reference patches have similar content.

For Unsupervised Learning, back-propagation loss, adversarial loss and flow-synthetic loss are optimized. If CycleGAN structure is included, cycle-consistency loss and identity loss are also optimized.

4.6.1 UnFlowSRNet

The pipeline overview of UnFlowSRNet can be seen in Figure 4.12.

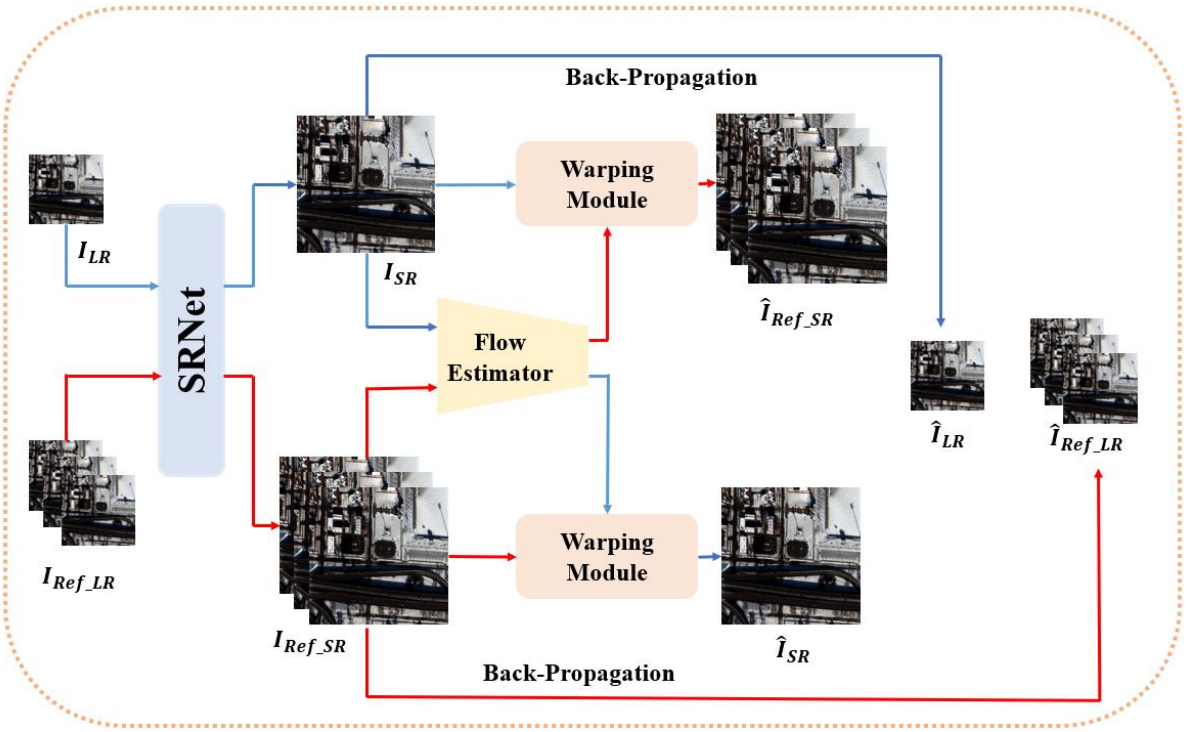
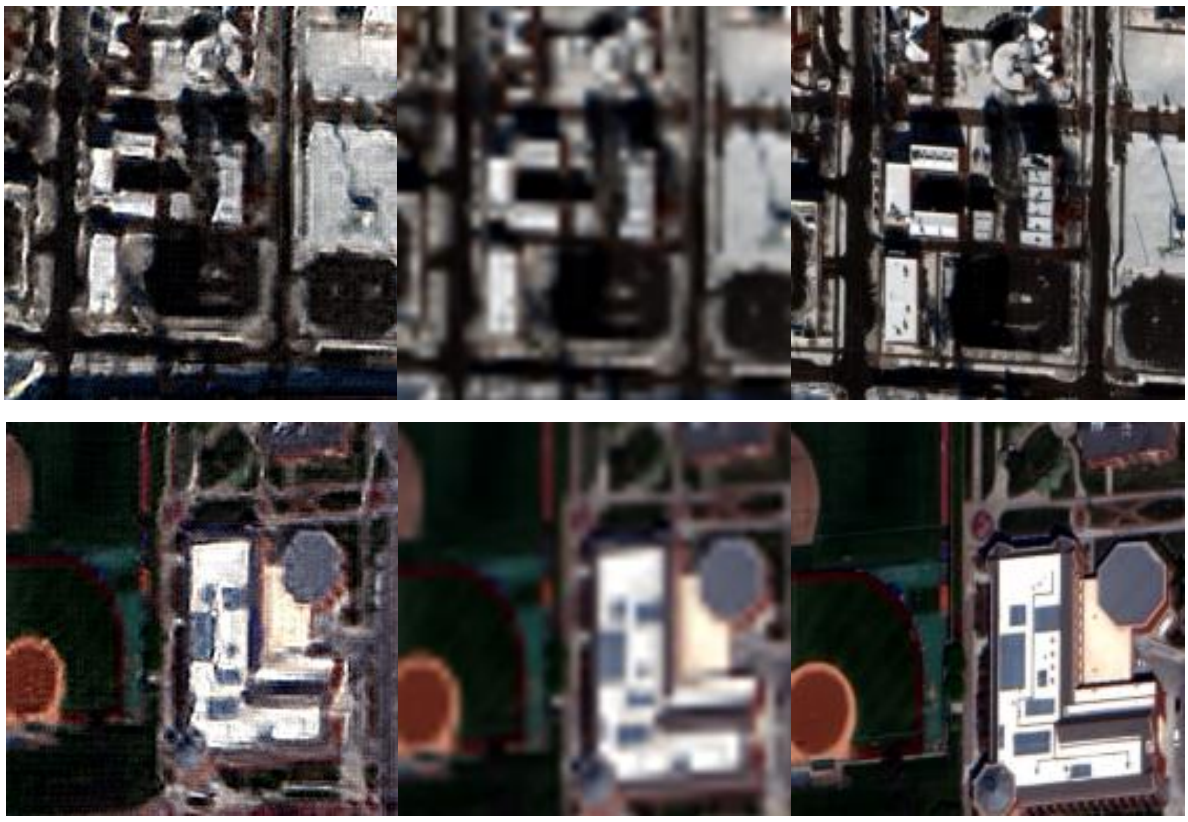


Figure 4.12: UnFlowSRNet

Since there is no ground truth, back-propagation loss is minimized to ensure data fidelity. Despite including adversarial learning, unsupervised learning with only super resolution network and optical flow estimation could not work in a satisfactory way. The PSNR performance of UnFlowSRNet (20.157) could be worse than simple bicubic interpolation (22.068) as shown in section 4.9. Example super resolution results are shown in figure 4.13.

The main reason for a worse performance of PSNR score of UnFlowSRNet than naïve bicubic interpolation is that there is no cross-scale information transformation between low resolution and high-resolution domain. Back-propagation loss could only measure pixel values in low-resolution space, while optical flow estimation is meant for multi-view feature transformation, which is based on the same scale. Only adversarial loss is meant for high resolution space. Thus, the super-resolved images have plenty of artificial textures. Another reason is that the training of UnFlowSRNet is on the original scale while testing is on the degraded scale.

One positive observation is that the super-resolved outputs of UnFlowSRNet have more finer textures than bicubic interpolation, which looks blurrier. Finer texture generation is essential for high-quality super resolution outputs.



(a) UnFlowSRNet (b) Bicubic (c) Ground Truth

Figure 4.13: Performance of UnFlowSRNet on Degraded Scale

4.6.2 UnFlowSRCycleGNet

Performance improvement could be achieved if CycleGAN structure is included. We then propose unsupervised optical-flow-based cycle-consistent image super resolution network (UnFlowSRCycleGNet) to generate super-resolved outputs without scaling. The pipeline is the same as seen in Figure 3.1. Structure of CycleGAN for image super resolution is shown in Figure 4.14.

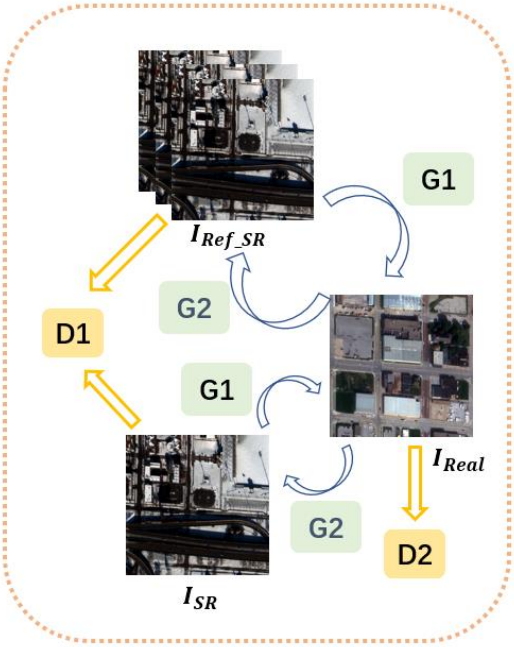
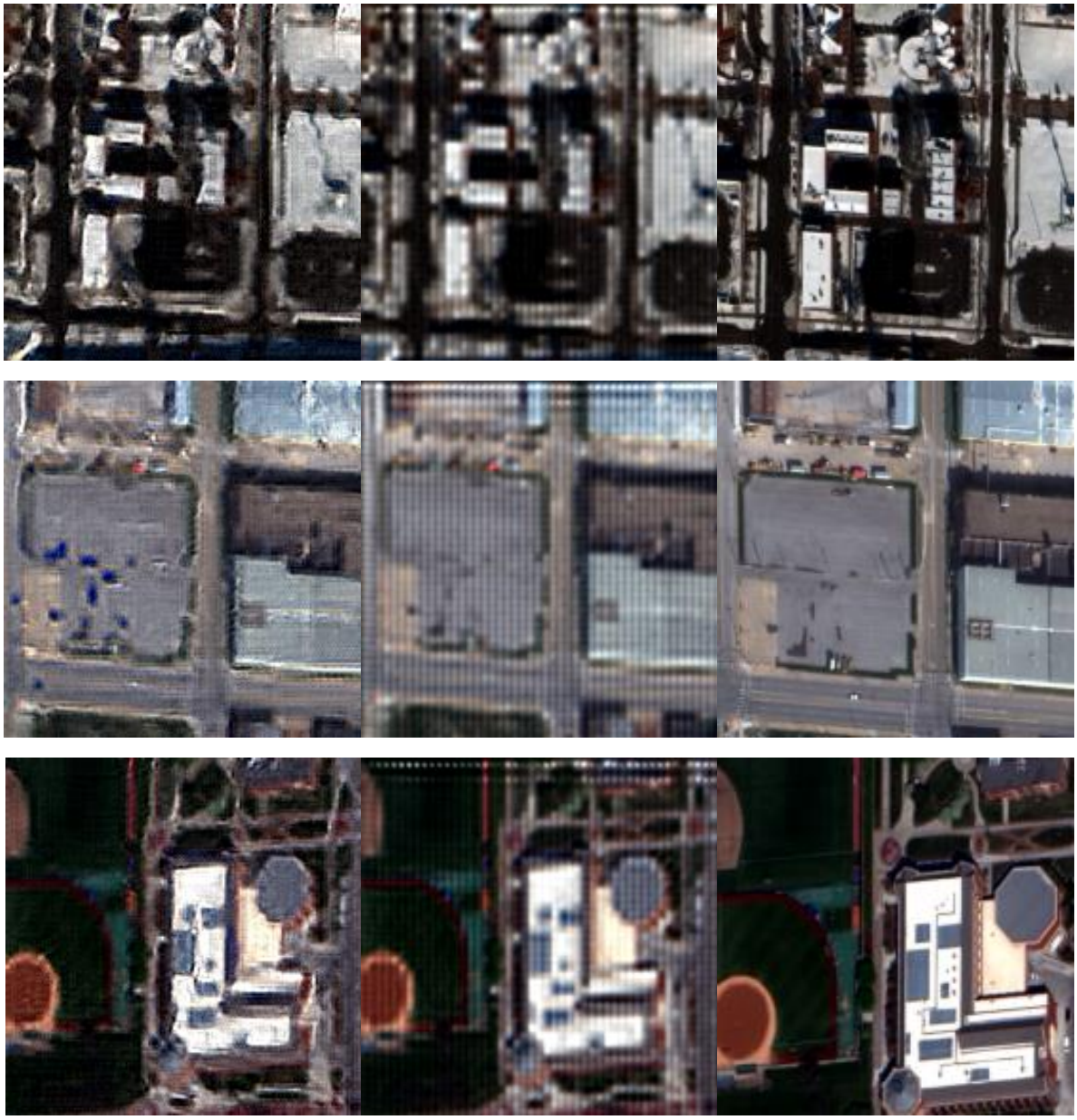


Figure 4.14: Super Resolution CycleGAN

Since GAN related losses are defined in the high-resolution domain for pure unsupervised learning without scaling, a refined GAN structure could be a way out. Although normally CycleGAN is utilized in style transformation rather than generating high-quality textures in super-resolved images, as we compare super-resolved outputs from UnFlowSRNet and ground truth, we humans could easily distinguish these two sets of images without labels, meaning that there is a hidden pattern or style difference between these two sets of images. Testing results show performance improvement of both PSNR and SSIM if CycleGAN is added as seen in section 4.9. Results in Figure 4.15 show that outputs from UnFlowSRCycleGNet have less artificial textures than UnFlowSRNet.

However, CycleGAN structure still could not settle the essential problem of cross-scale information transformation. The network fails to generate high-quality fine textures in super-resolved outputs. For style transformation, images with different style could both have rich high frequency information and transformation of patterns is different from generating unknown high frequency features. The CycleGAN structure serves more like icing on the cake to the network.



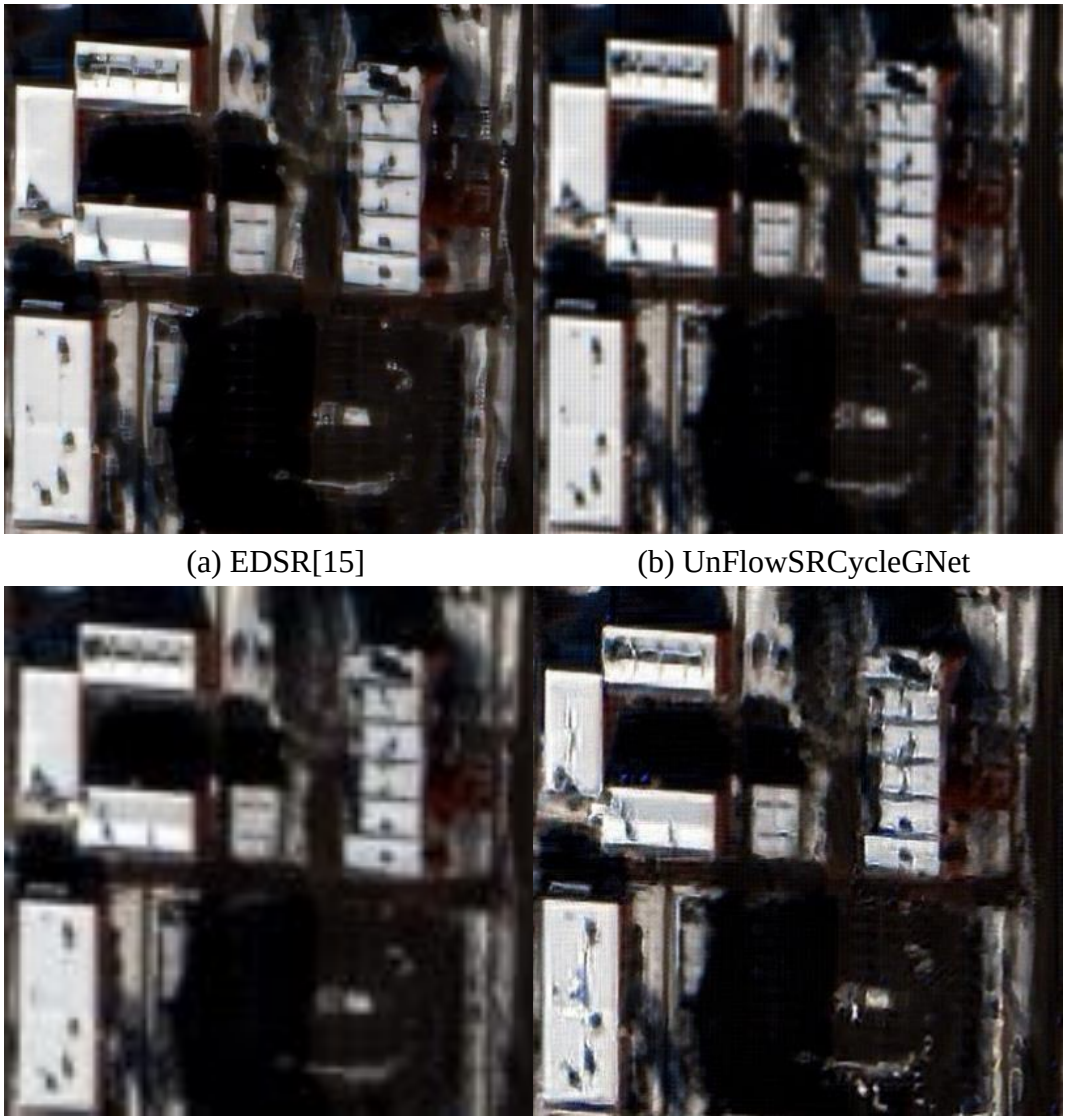
(a) UnFlowSRNet (b) UnFlowSRCycleGNet (c) Ground Truth

Figure 4.15: Performance Comparison between UnFlowSRNet and UnFlowSRCycleGNet on Degraded Scale

4.6.3 Testing on the Original Scale

One reason for the suboptimal performance of UnFlowSRNet and UnFlowSRCycleGNet might be that the testing scale is the wrong scale. Our unsupervised network is trained on the original scale rather than the degraded one. Testing on the original scale might show something different. Figure 4.16 shows super resolution outputs tested on the original scale.

As can be seen, CycleGAN structure could reduce artificial textures but could not produce adequate high-quality finer textures.



(a) EDSR[15] (b) UnFlowSRCycleGNet

(c) Bicubic (d) UnFlowSRNet

Figure 4.16: Unsupervised Learning Methods Tested on the Original Scale

4.7 Supervised Learning

Solly generating high-quality texture features in high-resolution space through unsupervised learning is difficult because the network fails to learn cross-scale features efficiently. Two solutions are proposed to address this problem: a) add external LR-HR image pairs and b) add feature extraction sharing mechanism to the super resolution module. Both methods involve supervised learning.

4.7.1 Add External LR-HR Image Pairs

One solution to learn cross-scale features for super resolution is to add external LR-HR image pairs[2]. For the training data, there is no ground truth, however, for the super resolution module, it can be trained in a supervised way with the help of the external image pairs. The network structure is the same as that in Figure 3.1, the only difference is that during training, there is a self-consistency loss for the super resolution module because of the external image pairs.

External LR-HR image pairs could be obtained by down-scaling image patches meant for adversarial training. Other settings of training are the same as unsupervised learning methods. Experiments show great performance improvement as seen in Figure 4.17. Quantitative performance is shown in section 4.9.

(a) UnFlowSRCycleGNet + External LR-HR Image Pairs

(b) UnFlowSRCycleGNet

Figure 4.17: Performance Comparison when External LR-HR Image Pairs are Included

4.7.2 FlowCircleSRCycleGNet

The problem for existing SISR methods is that their models are trained on the degraded scale which fails to meet our requirement. For these methods, the input image is down-scaled to generate its corresponding low-resolution image. In this way, the up-scaling process we learned is from the degraded scale to the original scale, rather than from the original scale to the up-sampled scale we desire. Learning on degraded scale is commonly used by SISR methods because scale-invariant features can also be learned from down-scaled images.

To guarantee that the super resolution process can also be applied to the scale we want, CircleSRNet is proposed to address this problem, where **feature extraction sharing mechanism** is introduced. The structure of CircleSRNet is shown in Figure 4.18.

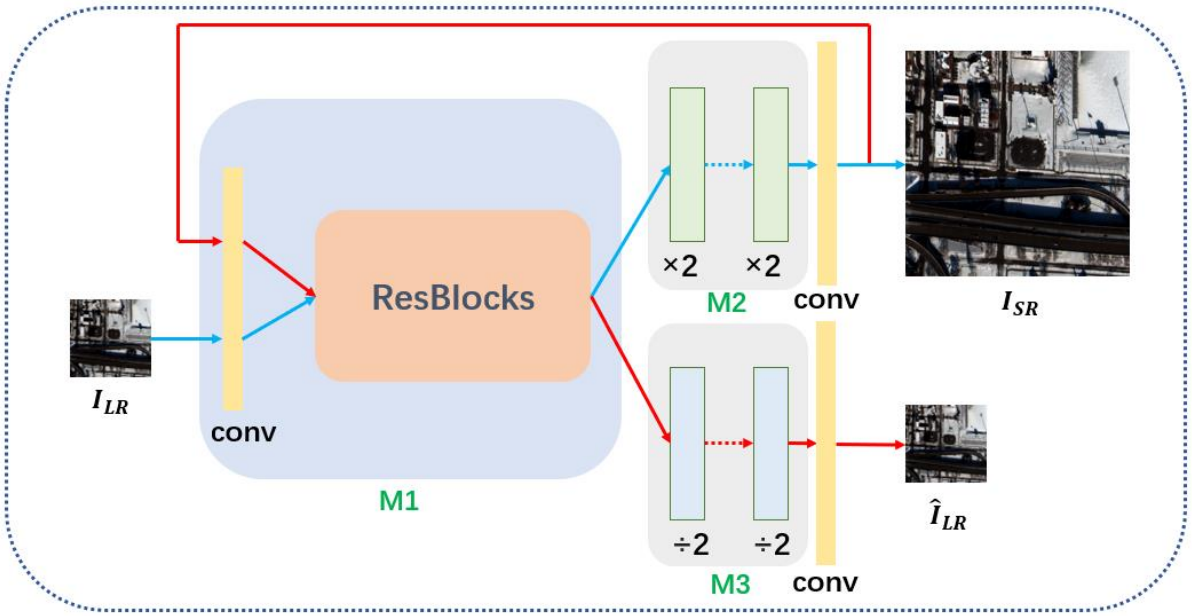


Figure 4.18: CircleSRNet Structure

Module M1 refers to the feature extraction part, M2 refers to the up-scaling process through pixel shuffle, and M3 refers to the down-scaling process through pixel un-shuffle, which is the reverse process to pixel shuffle. Up-scaling process can be shown in Figure 4.19, and down-scaling process can be shown in Figure 4.20.

Here, the image feature extraction part is shared for both up-scaling flow (the blue flow) and down-scaling flow (the red flow) as seen in Figure 4.18. First, low-resolution image is input to get its corresponding super-resolved output through up-scaling modules M1 and M2, then this super-resolved output is re-input into the network through down-scaling modules M1 and M3 to get its back-propagated low-resolution image. The feature extraction module M1 is shared in both processes, thus it could also learn features at the scale we want.

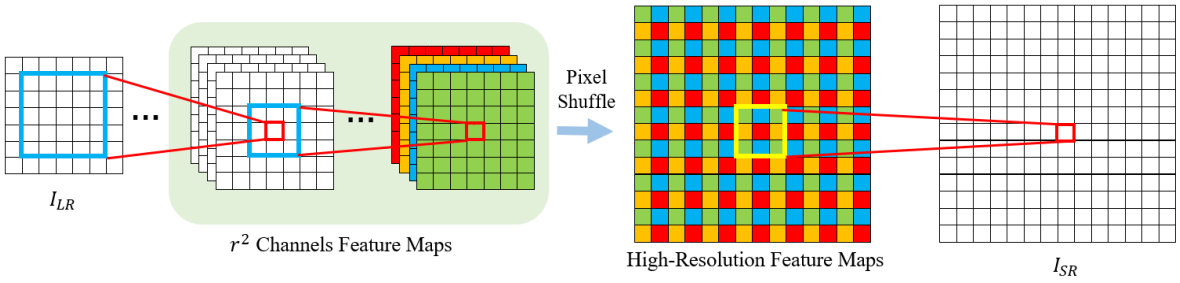


Figure 4.19: Up-Scaling Process of CircleSRNet with Pixel Shuffle[31]

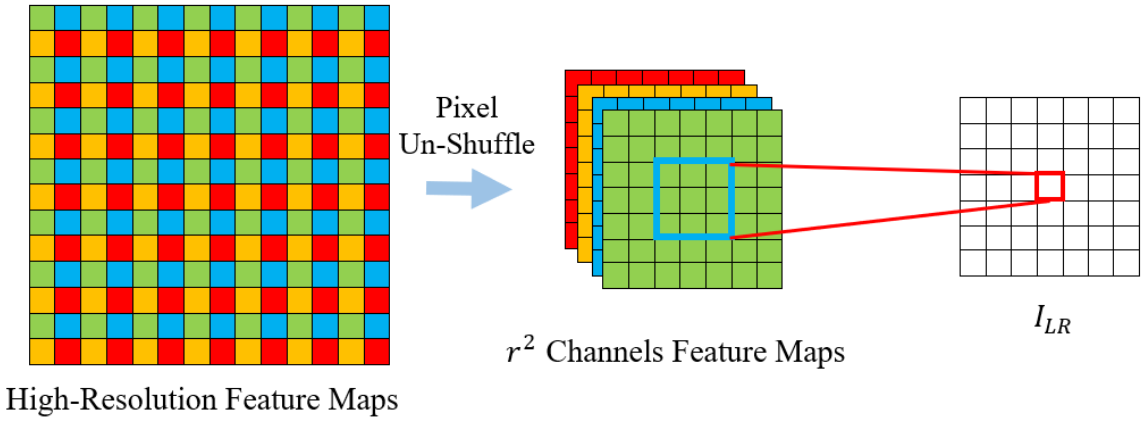


Figure 4.20: Down-Scaling Process of CircleSRNet with Pixel Un-Shuffle

For supervised learning involving down-scaling, 128×128 size patches are extracted and down-scaled to get corresponding LR-HR image pairs. Additional 128×128 size patches are randomly extracted from the same image for adversarial training. For references, patches with same size and same pixel location are extracted from random-shuffled neighboring images within same group to ensure the extracted reference patches have similar content.

According to the structure and information flow of the network, there are two losses that need to be minimized: a) self-consistency loss in the up-scaling process and b) back-propagation loss in the down-scaling process. Besides, perceptual loss, texture-matching loss, adversarial loss, flow-synthetic loss, cycle-consistency loss, identity loss and frequency loss are all optimized to guide the training of our network.

This CircleSRNet structure can be adopted as the super resolution module for our pipeline shown in Figure 3.1. The proposed model is FlowCircleSRCycleGNet. With the feature extraction sharing mechanism, we could train our model in a supervised way on the scale we want. Experiments show superior performance as can be seen in Figure 4.21. Quantitative analysis can be seen in section 4.9.

When tested on original scale, FlowCircleSRCycleGNet performed better than state-of-the-art SISR methods, as can be seen in Figure 4.22.

(a) UnFlowSRCycleGNet (b) FlowCircleSRCycleGNet (c) Ground Truth

Figure 4.21: Performance Comparison when CircleSRNet is Adopted on Degraded Scale

(a) Origin (b) Extracted Patch

(c) EDSR[15] (d) FlowCircleSRCycleGNet

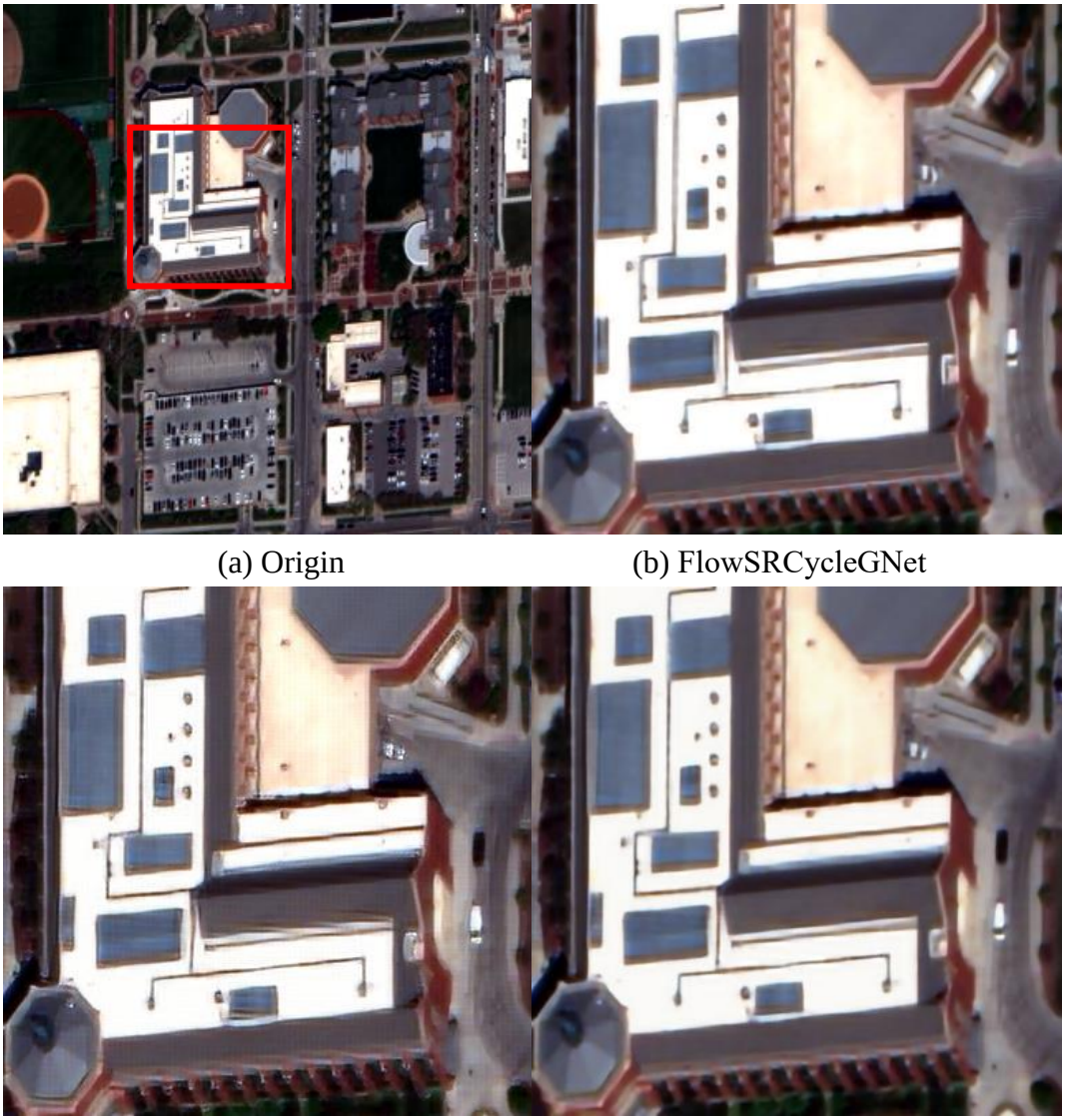
Figure 4.22: Performance of FlowCircleSRCycleGNet on Original Scale

4.7.3 Performance of Supervised Learning on Degraded Scale

If learning on degraded scale is allowed, then we can down-scale input patches to obtain LR-HR image pairs, where supervised learning with same pipeline as seen in Figure 3.1 can be conducted. We denote the model trained on degraded scale as FlowSRCycleGNet.

Self-consistency loss, perceptual loss, texture-matching loss, adversarial loss, flow-synthetic loss, cycle-consistency loss, identity loss and frequency loss are optimized to train FlowSRCycleGNet.

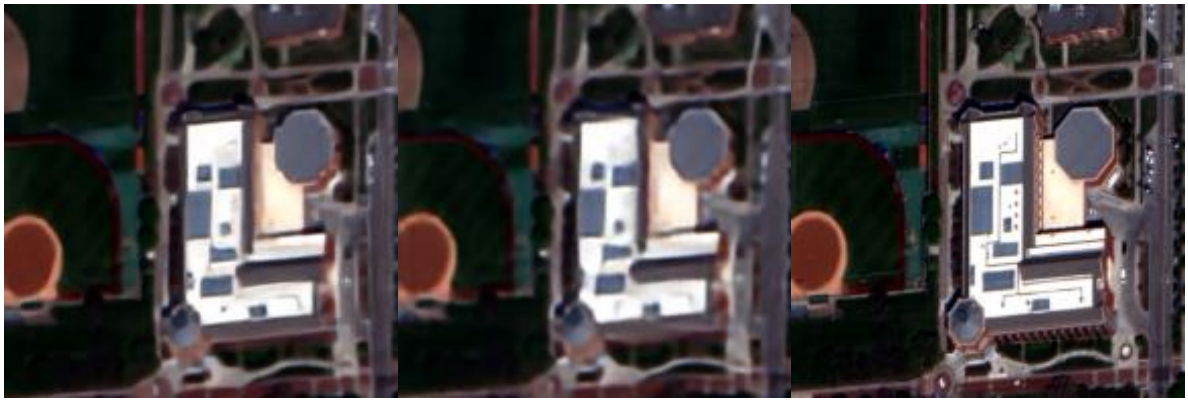
When compared with FlowCircleSRCycleGNet, although without feature extraction sharing mechanism, FlowSRCycleGNet could also generates high-quality super-resolved images as can be seen in Figure 4.23 and 4.24.



(a) Origin (b) FlowSRCycleGNet

(c) EDSR[15] (d) FlowCircleSRCycleGNet

Figure 4.23: Performance of FlowCircleSRCycleGNet Compared with FlowSRCycleGNet on the Original Scale

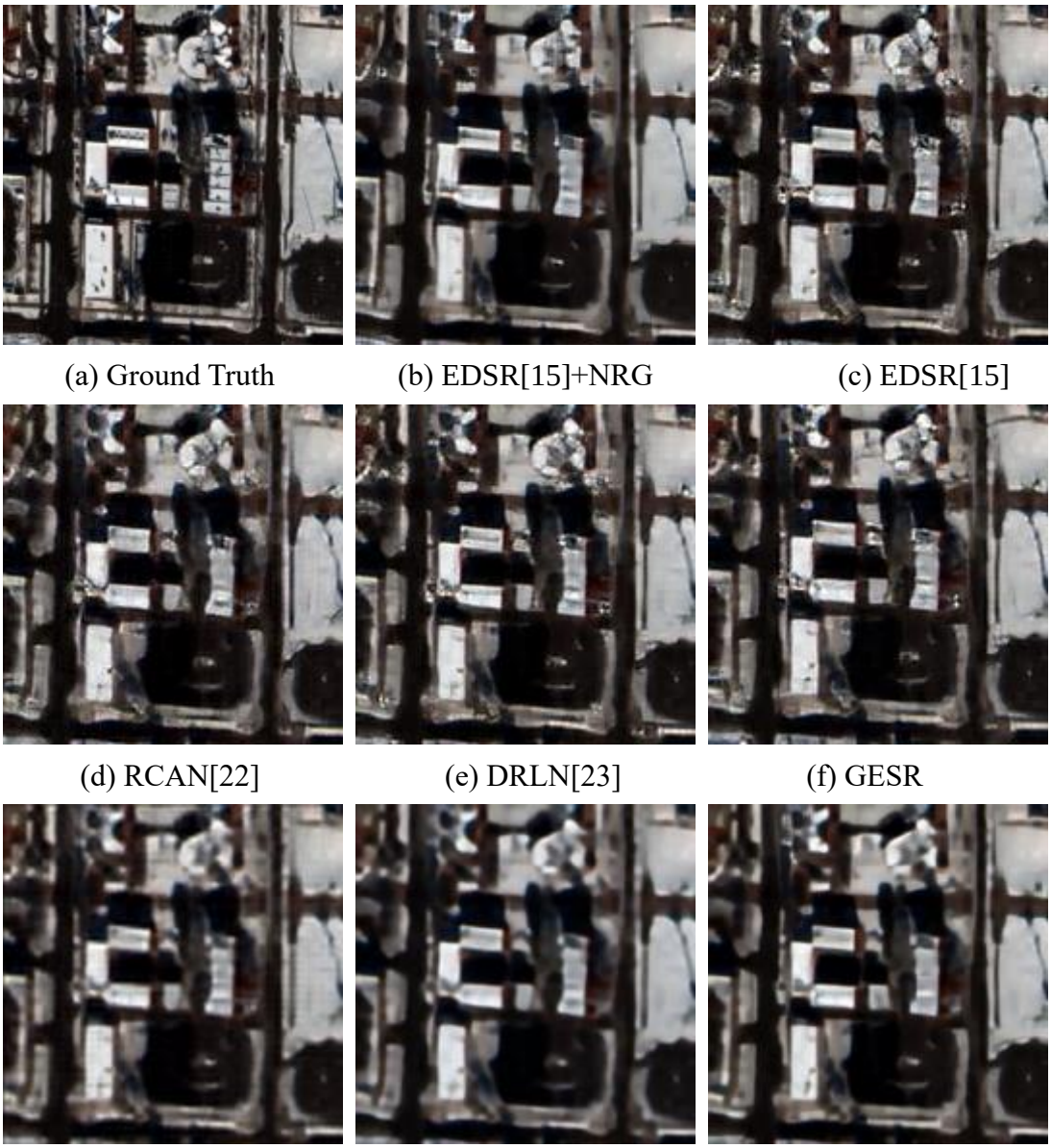


(a) FlowCircleSRCycleGNet (b) FlowSRCycleGNet (c) Ground Truth

Figure 4.24: Performance of FlowCircleSRCycleGNet Compared with FlowSRCycleGNet on Degraded Scale

4.8 Qualitative Evaluation Overview

Qualitative performance for supervised methods is shown in Figure 4.25. Model1 refers to UnFlowSRCycleGNet with external image pairs; Model2 refers to FlowCircleSRCycleGNet; Model3 refers to FlowSRCycleGNet.



(a) Ground Truth (b) EDSR[15]+NRG (c) EDSR[15]

(d) RCAN[22] (e) DRLN[23] (f) GESR

(g) Model1 (h) Model2 (i) Model3

Figure 4.25 Qualitative Performance for Supervised Methods

4.9 Quantitative Evaluation Overview

As shown in Table 4.2, With the help of Noise Residual Generator, EDSR[15] could perform better than other state-of-the-art methods in our dataset.

As shown in Table 4.3, pure unsupervised methods without scaling could not perform well in our pipeline, quantitative performance of UnFlowSRGNet and UnFlowSRCycleGNet are worse than simple bicubic interpolation. If external LR-HR image pairs are available, under the same pipeline as UnFlowSRCycleGNet, performance improves significantly and could be better than state-of-the-art SISR methods.

As shown in Table 4.4, FlowCircleSRCycleGNet could have superior performance than other SISR methods. Its performance is also comparable with FlowSRCycleGNet.

Models	**EDSR[15]**	**EDSR[15]+NRG**	**RCAN[22]**	**DRLN[23]**	**GESR**	**Bicubic**
PSNR	23.067	23.494	23.133	23.079	23.135	22.068
SSIM	0.7063	0.7329	0.7005	0.7114	0.7077	0.6268

Table 4.2: Performance of SISR Methods

Models	**UnFlowSRNet**	**UnFlowSRCycleGNet**	**UnFlowSRCycleGNet + External Image Pairs**
PSNR	20.157	21.867	23.6784
SSIM	0.5279	0.5990	0.7274

Table 4.3: Performance of Unsupervised Methods

Models	**FlowCircleSRCycleGNet**	**FlowSRCycleGNet**
PSNR	24.6426	24.5071
SSIM	0.7625	0.7708

Table 4.4: Performance of FlowCircleSRCycleGNet and FlowSRCycleGNet

5. Conclusion

5.1 Summary

We propose a super resolution network with optical flow estimation, feature extraction sharing mechanism and CycleGAN structure to tackle the super resolution problem where ground truth is unavailable and learning from original scale is required. Optical flow estimation could help the super resolution network to learn from multi-view information; feature extraction sharing mechanism could let our super resolution network to learn from initial scale rather than solely from degraded scale; CycleGAN structure could improve super-resolved output by translating styles from real high-resolution domain to super-resolved domain.

Unsupervised super resolution without scaling the input image is difficult because cross-scale feature information could not be learned efficiently. The task super resolution itself is a scale-changing process which require scale-invariant features from input image. The correspondence established by optical flow estimation between central image and references is on a stationary scale. Unlike typical reference-based super resolution task where high-resolution references are available, references in our situation are low-resolution images sharing similar content with the central image. Image super-resolved in this way would have plenty of artificial textures by adversarial training. Although CycleGAN structure could improve the performance, the outputs are still not high-quality super-resolved images with fine textures.

Two ways of building cross-scale information transformation are introduced: a) add external LR-HR image pairs and b) construct feature extraction sharing mechanism. Adding external LR-HR image pairs improve the performance. It has higher PSNR and SSIM score than state-of-the-art SISR methods. However, we could further improve the performance by implementing a feature extraction sharing mechanism. It is achieved by pixel Un-shuffle, which is the inverse process of pixel shuffle during super resolution. The pixel Un-shuffle process is used to down-scale the super-resolved image, so that our network could also learn features on original scale. Experiments have shown superior performance of this structure compared to other methods.

5.2 Future Work

The performance of supervised learning from degraded images can also be quite good on both degraded scale and original scale. No strong evidence shows that feature extraction sharing structure could outperform direct supervised learning on down-scaled inputs. Satellite images also comply with Image Internal Statistics, insist on learning from a certain scale is not convincing since super resolution itself is a scale changing process.

For unsupervised learning, we rely on the performance of generative adversarial networks, however, generative adversarial networks are hard to train and artificial textures are always presented in super-resolved outputs. More GAN structures apart from CycleGAN should be tried, as well as possible unsupervised methods other than GAN.

For learning correspondence from reference images, we adopt optical flow estimation. However, since brightness consistency constraint is not always conformed for satellite images, synthetic images tend to have different textures than ground truth. Maybe traditional feature analysis methods could be implemented as well. For example, Neural Texture Transfer[26] adopted a feature exchanging process between super-resolved output and references where alignment is no longer a requirement. The exchanging is conducted based on local texture similarity score.

Bibliography

[1] Anwar S, Khan S, Barnes N. A deep journey into super-resolution: A survey[J]. arXiv preprint arXiv:1904.07523, 2019.

[2] Yuan Y, Liu S, Zhang J, et al. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018: 701-710.

[3] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[4] Glasner D, Bagon S, Irani M. Super-resolution from a single image[C]//2009 IEEE 12th international conference on computer vision. IEEE, 2009: 349-356.

[5] Shaham T R, Dekel T, Michaeli T. Singan: Learning a generative model from a single natural image[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 4570-4580.

[6] Huang X, Li Y, Poursaeed O, et al. Stacked generative adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5077-5086.

[7] Shocher A, Cohen N, Irani M. “zero-shot” super-resolution using deep internal learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3118-3126.

[8] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

[9] Zontak M, Irani M. Internal statistics of a single natural image[C]//CVPR 2011. IEEE, 2011: 977-984.

[10] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[12] Sajjadi M S M, Scholkopf B, Hirsch M. Enhancenet: Single image super-resolution through automated texture synthesis[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4491-4500.

[13] Ulyanov D, Vedaldi A, Lempitsky V. Deep image prior[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9446-9454.

[14] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution[C]//European conference on computer vision. Springer, Cham, 2016: 694-711.

[15] Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017: 136-144.

[16] Zheng H, Ji M, Wang H, et al. CrossNet: An end-to-end reference-based super resolution network using cross-scale warping[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 88-104.
[17] Dong C, Loy C C, He K, et al. Image super-resolution using deep convolutional networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(2): 295-307.
[18] Yang J, Wright J, Huang T, et al. Image super-resolution as sparse representation of raw image patches[C]//2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008: 1-8.
[19] Timofte R, De Smet V, Van Gool L. Anchored neighborhood regression for fast example-based super-resolution[C]//Proceedings of the IEEE international conference on computer vision. 2013: 1920-1927.
[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
[22] Zhang Y, Li K, Li K, et al. Image super-resolution using very deep residual channel attention networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 286-301.
[23] Anwar S, Barnes N. Densely Residual Laplacian Super-Resolution[J]. arXiv preprint arXiv:1906.12021, 2019.
[24] Zheng H, Ji M, Han L, et al. Learning Cross-scale Correspondence and Patch-based Synthesis for Reference-based Super-Resolution[C]//BMVC. 2017.
[25] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
[26] Zhang Z, Wang Z, Lin Z, et al. Image super-resolution by neural texture transfer[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7982-7991.
[27] Dosovitskiy A, Fischer P, Ilg E, et al. Flownet: Learning optical flow with convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2758-2766.
[28] Sun D, Yang X, Liu M Y, et al. PWC-Net: CNNs for Optical Flow Using Pyramid[J]. Warping, and Cost Volume. arXiv. org, 2017.
[29] Garg R, BG V K, Carneiro G, et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]//European Conference on Computer Vision. Springer, Cham, 2016: 740-756.
[30] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.
[31] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1874-1883.

[32] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[C]//Advances in neural information processing systems. 2017: 5767-5777.

[33] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.